Theoretical Computer Science 416 (2012) 17–32

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

# On Müller context-free grammars☆

Zoltán Ésik, Szabolcs Iván *

*Department of Computer Science, University of Szeged, Hungary*

ARTICLE INFO

*Article history:*
Received 28 March 2011
Received in revised form 4 October 2011
Accepted 20 October 2011
Communicated by D. Perrin

*Keywords:*
Countable words
Context-free languages of countable words
Müller acceptance condition
Well-ordered
Scattered
Quasi-dense and dense words

ABSTRACT

We define context-free grammars with Müller acceptance condition that generate languages of countable words. We establish several elementary properties of the class of Müller context-free languages including closure properties and others. We show that every Müller context-free grammar can be transformed into a normal form grammar in polynomial space, and then we show that many decision problems can be decided in polynomial time for Müller context-free grammars in normal form. These decision problems include deciding whether the language generated by a normal form grammar contains only well-ordered, scattered, or dense words. In a further result, we establish a limitedness property of Müller context-free grammars: if the language generated by a grammar contains only scattered words, then either there is an integer $n$ such that each word of the language has Hausdorff rank at most $n$, or the language contains scattered words of arbitrarily large Hausdorff rank. We also show that it is decidable which of the two cases applies.



## 1. Introduction

In a general setting, a word over an alphabet $\Sigma$ is an isomorphism class of a linear order labeled in $\Sigma$. In this paper, we consider languages of countable words including scattered and dense words, i.e., words whose underlying linear order is scattered or dense; cf. [16].

Whereas finite automata over $\omega$-words and more generally countable and even uncountable words have been studied since the 1960s, cf. e.g., [6,7,1,17,18,2,5], context-free grammars generating countable words have received little attention.

Context-free grammars have been used to generate languages of $\omega$-words in [8,4,15]. Context-free grammars generating languages of countable words equipped with Büchi acceptance condition were considered in [10]. At the end of [10], we have also defined context-free grammars with Müller acceptance condition and showed that they generate a strictly larger class of languages. In this paper, our aim is to study Müller context-free languages in a systematic way.

We establish several elementary properties of the class of Müller context-free languages including closure properties and others. We show that every Müller context-free grammar can be transformed into a normal form grammar in polynomial space, and then we show that many decision problems can be solved in polynomial time for Müller context-free grammars in normal form. These decision problems include deciding whether the language generated by a normal form grammar contains only well-ordered, scattered, or dense words. In a further result, we establish a limitedness property of Müller context-free grammars: if the language generated by a grammar contains only scattered words, then either there is an integer $n$ such that each word of the language has Hausdorff rank at most $n$, or the language contains scattered words of arbitrarily large

☆ This research was supported by the National Foundation of Hungary for Scientific Research (OTKA) grant no. K75249, and by the project TÁMOP-4.2.1/B-09/1/KONV-2010-0005 "Creating the Center of Excellence at the University of Szeged", supported by the European Union and co-financed by the European Regional Fund.

* Corresponding author. Tel.: +36 62546396; fax: +36 62546397.
*E-mail address:* szabivan@inf.u-szeged.hu (S. Iván).

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.10.012

Hausdorff rank. We also prove that it is decidable which of the two cases applies and show that if the rank of the words in the language is bounded by some integer, then the least such bound is computable. Again, we give a polynomial time algorithm for grammars in normal form.

Countable words were first investigated in [9], where they were called "arrangements". Any countable word can be represented as the frontier of an infinite tree. Accordingly, any Müller context-free language can be seen as the frontier language of a tree language recognized by a Müller tree automaton; cf. [11]. Many of our decidability results can thus be deduced from certain closure properties and decidability results on Müller automata, but these arguments provide less information and higher complexity. See also Section 6.

## 2. Notation

In this section we recall some concepts for linear orders and words. A good reference on linear orders is [16].

A *partial order*, or partial ordering is a set $P$ equipped with a reflexive, transitive and antisymmetric binary relation, usually denoted $\leq$. We write $x < y$ if $x \leq y$ and $x \neq y$. A *linear order* is a partial order $(P, \leq)$ whose order relation is total, so that $x \leq y$ or $y \leq x$ for all $x, y \in P$. A countable (finite or infinite, respectively) linear order is a linear order which is a countable (finite or infinite, respectively) set. When $(P, \leq)$ and $(Q, \leq)$ are linear orders, an *isomorphism* (embedding, respectively) $(P, \leq) \to (Q, \leq)$ is a bijection (injection, respectively) $h\colon P \to Q$ such that $x \leq y$ implies $h(x) \leq h(y)$ for all $x, y \in P$. Two linear orders are *isomorphic* if there exists an isomorphism between them. In this case we also say that they have the same *order type*.

Below when there is no danger of confusion, we will denote a linear order just by $P, Q, \ldots$. Suppose that $P$ is a linear order. Then any subset $X$ of $P$ determines a *sub-order* of $P$ whose order relation is the restriction of the order relation of $P$ to $X$. Note that the inclusion function $X \hookrightarrow P$ is an embedding of $X$ into $P$. When in addition $X$ is such that for all $x, y \in X$ and $z \in P$, $x < z < y$ implies that $z \in X$, then we call $X$ an *interval*.

A linear order $(P, \leq)$ is a *well-order* if each nonempty subset of $P$ has a least element, and *dense* if it has at least two elements and for any $x < y$ in $P$ there is some $z$ with $x < z < y$.[1] A *quasi-dense* linear order is a linear order $(P, \leq)$ containing a dense linear sub-order, so that $P$ has a subset $P'$ such that $(P', \leq)$ is a dense order. A *scattered* linear order is a linear order which is not quasi-dense.

It is clear that every finite linear order is a well-order, every well-order is a scattered order, and every dense order is quasi-dense. It is well-known that up to isomorphism there are 4 countable dense linear orders: the rationals $\mathbb{Q}$ with the usual order and possibly endowed with either a least or a greatest element (or both).

An *ordinal* is an order type of a well-order. The finite ordinals $n$ are the order types of the finite linear orders. As usual, we denote by $\omega$ the least infinite ordinal which is the order type of the finite ordinals and the order type of the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$ equipped with the usual order. The order type of $\mathbb{Q}$ will be denoted $\eta$.

When $\tau$ and $\tau'$ are order types, we say that $\tau \leq \tau'$ if there is an embedding of a linear order of order type $\tau$ into a linear order of order type $\tau'$. The relation $\leq$ defined above is a linear order of the set of the countable ordinals.

We define several operations on linear orders. First, the reverse $P^r = (P, \leq^r)$ of a linear order $(P, \leq)$ is defined by $x \leq^r y$ if and only if $y \leq x$, for all $x, y \in P$. It is clear that the reverse of a scattered linear order is scattered, and the reverse of a dense linear order is dense.

Suppose that $P$ and $Q$ are linear orders. Then the *sum* $P + Q$ is the linear order on the disjoint union of $P$ and $Q$ such that $P$ and $Q$ are intervals of $P + Q$ and $x \leq y$ holds for all $x \in P$ and $y \in Q$. There is a more general notion. Suppose that $I$ is a linear order and for each $i \in I$, $P_i$ is a linear order. Then the generalized sum $P = \sum_{i \in I} P_i$ is obtained by replacing each point $i$ of $I$ with a copy of the linear order $P_i$. Formally, the generalized sum $P$ is the linear order on the disjoint union $\biguplus_{i \in I} P_i$ equipped with the order relation such that each $P_i$ is an interval and for all $i, j \in I$ with $i < j$, if $x \in P_i$ and $y \in P_j$ then $x < y$. The generalized sum gives rise to a *product* operation. Let $P$ and $Q$ be linear orders, and for each $y \in Q$, let $P_y$ be an isomorphic copy of $P$. Then $P \times Q$ is defined as the linear order $\sum_{y \in Q} P_y$. Note that this linear order is isomorphic to the linear order on the cartesian product of $P$ and $Q$ equipped with the order relation $(x, y) \leq (x', y')$ if and only if ($y < y'$ or ($y = y'$ and $x \leq x'$)).

**Lemma 1** ([16]). *Any scattered generalized sum of scattered linear orders is scattered. Similarly, any well-ordered generalized sum of well-orders is a well-order. Every quasi-dense linear order is a dense generalized sum of (nonempty) scattered linear orders.*

Thus, when $I$ is a scattered linear order and for each $i \in I$, $P_i$ is a scattered linear order, then so is $\sum_{i \in I} P_i$, and similarly for well-orders. And if $P$ is a quasi-dense linear order, then there is a dense linear order $D$ and (nonempty) scattered linear orders $P_x$, $x \in D$ such that $P$ is isomorphic to $\sum_{x \in D} P_x$.

The above operations preserve isomorphism, so that they give rise to corresponding operations $\tau + \tau'$ and $\tau \times \tau'$ on order types. In particular, the sum and product of two ordinals is well-defined (and is an ordinal). The reverse of an order type $\tau$ will be denoted $-\tau$. The ordinals are also equipped with the exponentiation operation; cf. [16].

An *alphabet* is a finite nonempty set $\Sigma$. A *word* over an alphabet $\Sigma$ is a labeled linear order, i.e., a triple $u = (\mathrm{dom}(u), \leq_u, \lambda_u)$, where $(\mathrm{dom}(u), \leq_u)$ is a linear order and $\lambda_u$ is a labeling function $\mathrm{dom}(u) \to \Sigma$. The underlying linear order $\mathrm{dom}(\varepsilon)$ of

[1] In [16], a singleton linear order is also called dense.

the empty word $\varepsilon$ is the empty linear order. We say that a word is *finite* (*infinite* or *countable*, respectively), if its underlying linear order is finite (infinite or countable, respectively). An *isomorphism* of words is an isomorphism of the underlying linear orders that preserves the labeling. *Embeddings* of words are defined in the same way. We usually identify isomorphic words. We will say that a word $u$ is a *subword* of a word $v$ if there is an embedding $u \hookrightarrow v$. When in addition the image of the underlying linear order of $u$ is an interval of the underlying linear order of $v$ we call $u$ a *factor* of $v$.

The *order type of a word* is the order type of its underlying linear order. Thus, the order type of a finite word is a finite ordinal. A word whose order type is $\omega$ is called an $\omega$-*word*.

The *reverse* of a word $u = (P, \leq, \lambda)$ is $u^r = (P, \leq^r, \lambda)$, where $(P, \leq^r)$ is the reverse of $(P, \leq)$. Suppose that $u = (P, \leq_u, \lambda_u)$ and $v = (Q, \leq_v, \lambda_v)$ are words over $\Sigma$. Then their *concatenation* $uv$ is the word over $\Sigma$ whose underlying linear order is $P + Q$ and whose labeling function agrees with $\lambda_u$ on points in $P$, and with $\lambda_v$ on points in $Q$. More generally, when $I$ is a linear order and $u_i$ is a word over $\Sigma$ with underlying linear order $P_i = (\mathrm{dom}(u_i), \leq_i)$, for each $i \in I$, then the *generalized concatenation* $\prod_{i\in I} u_i$ is the word whose underlying linear order is $\sum_{i\in I} P_i$ and whose labeling function agrees with the labeling function of $P_i$ on the elements of each $P_i$. In particular, when $u_0, u_1, \dots, u_n, \dots$ are words over $\Sigma$ and $I$ is the linear order $\omega$ or its reverse $-\omega$, then $\prod_{i\in I} u_i$ is the word $u_0u_1 \dots u_n \dots$ or $\dots u_n \dots u_1u_0$, respectively. When $u_i = u$ for each $i$, these words are denoted $u^\omega$ and $u^{-\omega}$, respectively.

Some examples of words over the alphabet $\Sigma = \{a, b\}$ are the finite word $aab$ which is the (isomorphism class of the) 3-element labeled linear order $\{0 < 1 < 2\}$ whose points are labeled $a$, $a$ and $b$, in this order. Examples of infinite words are $a^\omega$ and $a^{-\omega}$, whose order types are $\omega$ and $-\omega$, respectively, such that each point is labeled $a$. For another example, consider the linear order $\mathbb{Q}$ of the rationals and label each point $a$. The resulting word of order type $\eta$ is denoted $a^\eta$. More generally, let $\Sigma$ contain the (different) letters $a_1, \dots, a_n$. Then up to isomorphism there is a unique labeling of the rationals such that between any two points there are $n$ points labeled $a_1, \dots, a_n$, respectively. The resulting word is denoted $(a_1, \dots, a_n)^\eta$; cf. [12].

In the sequel, we will also make use of the *substitution* operation on words. Suppose that $u$ is a word over $\Sigma$ and for each letter $a \in \Sigma$, $u_a$ is a word over $\Delta$. Then the word $u[a \leftarrow u_a]_{a\in\Sigma}$ obtained by substituting $u_a$ for each occurrence of a letter $a$ in $u$ (or replacing each occurrence of a letter $a$ with $u_a$) is formally defined as follows. Let $u = (P, \leq, \lambda)$ and $u_a = (P_a, \leq_a, \lambda_a)$ for each $a \in \Sigma$. Then for each $i \in P$ let $u_i = (P_i, \leq_i, \lambda_i)$ be an isomorphic copy of $u_{\lambda(i)}$. We define

$$u[a \leftarrow u_a]_{a\in\Sigma} = \prod_{i\in P} u_i.$$

Note that when $u = a^\omega$, then $u[a \leftarrow v]$ is $v^\omega$, and similarly for $v^{-\omega}$.

For any words $u_1, \dots, u_n$ over an alphabet $\Sigma$, we define

$$(u_1, \dots, u_n)^\eta = (a_1, \dots, a_n)^\eta[a_1 \leftarrow u_1, \dots, a_n \leftarrow u_n].$$

We call a word over an alphabet $\Sigma$ *well-ordered*, *scattered*, *dense*, or *quasi-dense* if its underlying linear order has the appropriate property. For example, the words $a^\omega$, $a^\omega b^\omega a$, $(a^\omega)^\omega$ over the alphabet $\{a, b\}$ are well-ordered, the words $a^\omega a^{-\omega}$, $a^{-\omega}a^\omega$ are scattered but not well-ordered, the words $a^\eta$, $a^\eta b a^\eta$, $(a, b)^\eta$ are dense, and the words $(ab)^\eta$, $(a^\omega)^\eta$, $(ba^\eta b)^\omega$ are quasi-dense but not dense.

From Lemma 1 we immediately have:

**Lemma 2.** *Any scattered generalized concatenation of scattered words is scattered. Any well-ordered generalized concatenation of well-ordered words is well-ordered. Moreover, every quasi-dense word is a dense concatenation of (nonempty) scattered words.*

As already mentioned, we will usually identify isomorphic words, so that a word is an isomorphism type (or isomorphism class) of a labeled linear order. When $\Sigma$ is an alphabet, we let $\Sigma^*$, $\Sigma^\omega$ and $\Sigma^\infty$ respectively denote the set of all finite words, $\omega$-words, and *countable* words over $\Sigma$. We let $\Sigma^+$ and $\Sigma^{+\infty} = \Sigma^\infty\Sigma\Sigma^\infty$ respectively denote the set of all finite nonempty words and the set of all countable nonempty words over $\Sigma$. The length of a finite word $w$ will be denoted $|w|$. As an extension, we let $\mathbb{N}^*$ ($\mathbb{N}^\omega$, respectively) stand for the set of all sequences of the form $n_1 \dots n_k$ ($n_1n_2 \dots$, respectively) with each $n_i$ being in $\mathbb{N}$. (Note that $\mathbb{N}$ is not an alphabet since it is infinite.) We also define two partial orders as follows. The *prefix order* $x \leq_{\mathrm{pr}} y$ holds for $x, y \in \mathbb{N}^*$ if and only if $y = xx'$ for some $x' \in \mathbb{N}^*$, and the *lexicographic order* $x \leq_{\mathrm{lex}} y$ holds for $x, y \in \mathbb{N}^* \cup \mathbb{N}^\omega$ if and only if $x = y$ or $x = wix'$ and $y = wjy'$ for some $w \in \mathbb{N}^*$, $x', y' \in \mathbb{N}^* \cup \mathbb{N}^\omega$ and $i, j \in \mathbb{N}$ with $i < j$. It is clear that any subset of $\mathbb{N}^*$ containing pairwise incomparable elements with respect to the prefix order is linearly ordered by the lexicographic order.

A *language over* $\Sigma$ is any subset $L$ of $\Sigma^\infty$. When $L \subseteq \Sigma^*$ or $L \subseteq \Sigma^\omega$, we sometimes call $L$ a *language of finite words* or $\omega$-*words*, or an $\omega$-*language*.

Languages are equipped with several operations. First of all, they are equipped with the usual set theoretic operations. We now define the generic operation of language substitution.

Suppose that $u \in \Sigma^\infty$ and for each $a \in \Sigma$, $L_a \subseteq \Delta^\infty$. Then the words in the language $u[a \leftarrow L_a]_{a\in\Sigma} \subseteq \Delta^\infty$ are obtained from $u$ by substituting in all possible ways a word in $L_a$ for each occurrence of each letter $a \in \Sigma$. Different occurrences of the same letter $a$ may be replaced by different words in $L_a$.

Formally, suppose that $u = (P, \leq, \lambda)$. For each $x \in P$ with $\lambda(x) = a$, let us choose a word $u_x = (P_x, \leq_x, \lambda_x)$ which is isomorphic to some word in $L_a$. Then the language $u[a \leftarrow L_a]_{a\in\Sigma}$ consists of all words $\prod_{x\in P} u_x$.

Suppose now that $L \subseteq \Sigma^\infty$ and for each $a \in \Sigma$, $L_a \subseteq \Delta^\infty$. Then

$$L[a \leftarrow L_a]_{a\in\Sigma} = \bigcup_{u\in L} u[a \leftarrow L_a]_{a\in\Sigma}.$$

We call $L[a \leftarrow L_a]_{a\in\Sigma}$ the language *obtained from L by substituting the language $L_a$ for each $a \in \Sigma$*.

As mentioned above, set theoretic operations on languages in $\Sigma^\infty$ have their standard meaning. Below we define some other operations. Let $L, L_1, L_2, \ldots, L_m \subseteq \Sigma^\infty$. Then we define:

1. $L_1L_2 = \{ab\}[a \leftarrow L_1, b \leftarrow L_2] = \{uv : u \in L_1, v \in L_2\}$.
2. $L^* = \{a\}^*[a \leftarrow L] = \{u_1 \ldots u_n : n < \omega,\ u_i \in L\}$.
3. $L^\omega = \{a^\omega\}[a \leftarrow L] = \{u_0u_1 \ldots u_n \ldots : u_i \in L\}$.
4. $L^{-\omega} = \{a^{-\omega}\}[a \leftarrow L] = \{\ldots u_n \ldots u_1u_0 : u_i \in L\}$.
5. $(L_1, \ldots, L_m)^\eta = \{(a_1, \ldots, a_m)^\eta\}[a_1 \leftarrow L_1, \ldots, a_m \leftarrow L_m]$.
6. $L^\infty = \{a\}^\infty[a \leftarrow L]$.

The above operations are respectively called *concatenation*, *star*, $^\omega$*-power*, $^{-\omega}$*-power*, $^\eta$*-power*, and $^\infty$*-power*.

Some more operations. The *reverse* $L^r$ of a language $L \subseteq \Sigma^\infty$ is defined as $L^r = \{u^r : u \in L\}$. The language $\mathrm{Pre}(L)$ is given by $\mathrm{Pre}(L) = \{u : \exists v\ uv \in L\}$ and $\mathrm{Suf}(L)$ is defined symmetrically. $\mathrm{In}(L)$ is $\{u : \exists v, w\ vuw \in L\}$, and $\mathrm{Sub}(L)$ is the collection of all words $u$ such that there is an embedding $u \hookrightarrow v$ for some $v \in L$.

When $A$ is a set, we denote by $P(A)$ the *power set of* $A$ and by $P_+(A)$ the set of all nonempty subsets of $A$.

## 3. Tree domains and Müller context-free grammars

A subset $X$ of $\mathbb{N}^*$ is called *prefix closed* if whenever $u \cdot v \in X$ for some words $u, v \in \mathbb{N}^*$, then also $u \in X$. A *tree domain* is an arbitrary nonempty, prefix closed subset $T$ of $\mathbb{N}^*$ (whose elements are usually referred as *nodes of* $T$) with each node $x \in T$ having a finite number of *successors*, i.e., $x \cdot i \in T$ for only finitely many $i \in \mathbb{N}$. The number of successors of a node $x \in T$ is called the *degree of* $x$ *in* $T$ and is denoted $\deg_T(x)$, or simply $\deg(x)$ when $T$ is understood. A node $u$ is a *descendant* of a node $v$ if $v \le_{\mathrm{pr}} u$. If in addition $u \neq v$, then $u$ is a *strict descendant* of $v$. Accordingly, $v$ is an *ancestor*, or a *strict ancestor* of $u$. A *path* of a tree domain is a nonempty prefix closed set $\pi \subseteq T$ such that for any $x \in \pi$, $\deg_\pi(x) \le 1$. A path can be either finite or infinite. When $\pi$ is a finite (infinite, respectively) path, we also write $\pi = x_0, \ldots, x_n$ ($\pi = x_0, x_1, \ldots$, respectively) to indicate that $\pi = \{x_i : 0 \le i \le n\}$ ($\pi = \{x_i : i \ge 0\}$, respectively) and for each $i \ge 0$, $x_{i+1}$ is a successor of $x_i$. When $\pi = \{x_0, x_1, \ldots\}$ is an infinite path, we sometimes identify $\pi$ with the (unique) word $u \in \mathbb{N}^\omega$ such that $\pi$ consists of the finite prefixes of $u$.

A path $\pi$ of $T$ is *maximal* if for any path $\pi'$ of $T$ with $\pi \subseteq \pi'$ we have $\pi = \pi'$. It is clear that for any node $x \in T$ there exists at least one maximal path of $T$ containing $x$, moreover, when $x$ and $y$ are pairwise incomparable nodes with respect to the prefix order, then no path can contain both of them. Nodes of a tree domain $T$ with degree 0 are called *leaves of* $T$, while nodes with a positive degree are called the *inner nodes of* $T$. When $T$ is a tree domain and $x \in T$ is a node of $T$, then the *sub-tree domain of* $T$ *rooted at* $x$ is the tree domain $T|_x = \{y \in \mathbb{N}^* : x \cdot y \in T\}$.

A *context-free grammar with Müller acceptance condition*, or MCFG for short, is a system $G = (V, \Sigma, P, S, \mathcal{F})$ where $V$ is the finite, nonempty set of *nonterminals* (or variables), $\Sigma$ is the *terminal* alphabet with $V \cap \Sigma = \emptyset$, $S \in V$ is the *start symbol*, $P$ is the finite set of *productions* (or rules) of the form $A \to \alpha$ with $A \in V$ and $\alpha \in (V \cup \Sigma)^*$, and $\mathcal{F} \subseteq P_+(V)$ is the *Müller acceptance condition*.

A *derivation tree* of the above grammar $G$ is a mapping $t : \mathrm{dom}(t) \to \Sigma \cup V \cup \{\varepsilon\}$ where $T = \mathrm{dom}(t)$ is a tree domain satisfying the following conditions:

- Each inner node $x \in T$ is labeled by some nonterminal $A$, i.e., $t(x) = A$ for some $A \in V$;
- For any node $x \in T$ and $i \in \mathbb{N}$, $x \cdot i \in T$ if and only if $1 \le i \le \deg(x)$;
- For any node $x \in T$ with $\deg(x) > 0$ exactly one of the following conditions holds:
  1. either there exists a production $A \to X_1 \ldots X_k$ in $P$, where $A = t(x)$, $k = \deg(x)$ and for each $1 \le i \le k$, $X_i = t(x \cdot i)$ is a member of $V \cup \Sigma$;
  2. or $\deg(x) = 1$, $t(x \cdot 1) = \varepsilon$ and $A \to \varepsilon$ is in $P$, where $A = t(x)$.
- Finally, *t satisfies the Müller acceptance condition* $\mathcal{F}$: for each infinite (maximal) path $\pi$ of $T$ the set

  $$\mathrm{InfLab}_t(\pi) = \{A \in V : A = t(x) \text{ for infinitely many } x \in \pi\}$$

  is a member of $\mathcal{F}$.

When $x \in \mathrm{dom}(t)$ is a node of $t$, then the *subtree of t rooted at x* is the derivation tree $t|_x$ with $\mathrm{dom}(t|_x) = \mathrm{dom}(t)|_x$ and $t|_x(y)$ defined as $t(x \cdot y)$, for each $y \in \mathrm{dom}(t|_x)$.

A derivation tree is *complete* if each of its leaves is labeled in $\Sigma \cup \{\varepsilon\}$. For any set $V' \subseteq V$ of nonterminals and symbol $X \in V' \cup \Sigma$, let $\Delta_G(V', X)$ denote the set of those derivation trees $t$ of $G$ having root symbol $X$ which are labeled in $V' \cup \Sigma \cup \{\varepsilon\}$, i.e. $t(\varepsilon) = X$ and for each node $x \in \mathrm{dom}(t)$, $t(x) \in V' \cup \Sigma \cup \{\varepsilon\}$. When $G$ is clear from the context, we omit the subscript.

*Nodes, leaves, etc.* of $t$ are defined as nodes, leaves etc. of its domain $\mathrm{dom}(t)$. The *frontier* of a derivation tree $t$ is the word $\mathrm{fr}(t) = (L, <, t')$ where $L \subseteq \mathrm{dom}(t)$ is the set of those leaves of $t$ which are labeled in $V \cup \Sigma$, $<$ is the restriction of the lexicographic order to $L$, and $t'$ is the restriction of $t$ onto $L$.

We write $X \Rightarrow_G^\infty \alpha$ for a symbol $X \in V \cup \Sigma$ and a word $\alpha \in (V \cup \Sigma)^\infty$ if there exists a derivation tree $t \in \Delta(V, X)$ with $\mathrm{fr}(t) = \alpha$. If there is a *finite* derivation tree with these properties, we write $X \Rightarrow_G^* \alpha$. When $G$ is clear from the context, we omit the subscripts. For a nonterminal $A \in V$, we denote by $L^\infty(G, A)$ the set $\{w \in \Sigma^\infty : A \Rightarrow^\infty w\}$. The *language generated by* $G$ is $L^\infty(G) = L^\infty(G, S)$.

**Definition 3.** A language $L \subseteq \Sigma^\infty$ is a *Müller context-free language*, or an MCFL if $L = L^\infty(G)$ for some MCFG $G = (V, \Sigma, P, S, \mathcal{F})$.

In [10], we studied *Büchi context-free languages*: a Büchi context-free grammar, or BCFG is a system $G = (V, \Sigma, P, S, F)$ where $V, \Sigma, P$ and $S$ are the same as in the case of an MCFG and $F \subseteq V$ is a Büchi acceptance condition. In this case, a derivation tree $t$ of $G$ has to satisfy the condition that $\mathrm{InfLab}(\pi) \cap F \neq \emptyset$ for each infinite path $\pi$ of $t$. The Büchi context-free language, or BCFL generated by the above BCFG $G$ is $L^\infty(G) = \{\mathrm{fr}(t) \in \Sigma^\infty : t \text{ is a derivation tree of } G\}$.

**Example 4.** The BCFG $G = (V, \Sigma, P, S, \{S\})$ where $P = \{S \to SS, S \to \varepsilon\} \cup \{S \to a : a \in \Sigma\}$ generates $\Sigma^\infty$, the set of all countable words.

It is clear that the class of BCFLs is contained in the class of MCFLs; in [10], it has been shown that the inclusion is strict. Below we give two further examples that are MCFLs but not BCFLs.

**Example 5.** Let $\Sigma$ be an alphabet and consider the MCFG

$$G = (\{S, I\}, \Sigma, P, S, \{\{I\}\}) \text{ where}$$
$$P = \{S \to a : a \in \Sigma \cup \{\varepsilon\}\} \cup \{S \to I\} \cup \{I \to SI\}.$$

Then $L^\infty(G)$ is the set of all countable well-ordered words over $\Sigma$.

To see this, we show that for each countable ordinal $\alpha$, the set of those words $u \in \Sigma^\infty$ having order type $\alpha$ is a subset of $L$. For $\alpha = 0$ and $\alpha = 1$ the statement holds, since $S \to \varepsilon$ and $S \to a$, $a \in \Sigma$ are productions of $G$. Assume the claim holds for each ordinal less than $\alpha$ and let $u \in \Sigma^\infty$ be a word having order type $\alpha$. Then $u = \prod_{i<\omega} u_i$ for some (possibly empty) words $u_i$, each having order type less than $\alpha$. Applying the induction hypothesis we get that $S \Rightarrow^\infty u_i$ for each $i < \omega$, and by $S \Rightarrow^\infty S^\omega$ we have $S \Rightarrow^\infty u$ proving the claim.

We do not show here that $L$ contains well-ordered words only: in Section 6 we give a decision procedure using which one can check whether an MCFL given by an MCFG consists of well-ordered words only.

**Example 6.** Let $\Sigma$ be an alphabet and consider the MCFG

$$G = (\{S, I^-, I^+\}, \Sigma, P, S, \{\{I^-\}, \{I^+\}\}) \text{ where}$$
$$P = \{S \to a : a \in \Sigma \cup \{\varepsilon\}\} \cup \{S \to I^- I^+\} \cup \{I^+ \to SI^+\} \cup \{I^- \to I^- S\}.$$

Then $L^\infty(G)$ is the set of all countable scattered words over $\Sigma$.

In [10] it has been shown that neither the set of all well-ordered words, nor the set of all scattered words is a BCFL.

## 4. Complexity of emptiness

In this section we show that the following question is complete for PSPACE: given an MCFG $G$, does it hold that $L^\infty(G) = \emptyset$? To show this, we use a corresponding complexity result for Müller tree automata.

A *rank type* is a finite nonempty set $\mathcal{R}$ of nonnegative integers. A *ranked alphabet* of rank type $\mathcal{R}$ is a disjoint union $\Delta = \biguplus_{k \in \mathcal{R}} \Delta_k$ where each $\Delta_k$ is a nonempty finite set of *$k$-ary symbols. A $\Delta$-tree* is a mapping $t : \mathrm{dom}(t) \to \Delta$ satisfying all the following conditions:

- $\mathrm{dom}(t)$ is a tree domain;
- whenever $t(u) \in \Delta_k$ for some integer $k$, then for each $i \in \mathbb{N}$, $u \cdot i \in \mathrm{dom}(t)$ if and only if $1 \leq i \leq k$.

A *$\Delta$-tree language* is an arbitrary set of $\Delta$-trees.

A (deterministic) *Müller tree automaton* is a tuple $M = (Q, \Delta, \delta, q_0, F_{\mathrm{fin}}, \mathcal{F})$, where $Q$ is the finite set of states, $\Delta$ is a ranked alphabet of some rank type $\mathcal{R}$, $\delta$, is a family $\{\delta_k : k \in \mathcal{R}\}$ of transition functions, each $\delta_k$ mapping $Q \times \Delta_k$ to $Q^k$, $q_0 \in Q$ is the *start state*, $F_{\mathrm{fin}} \subseteq Q \times \Delta_0$ is the set of *final configurations* and $\mathcal{F} \subseteq P_+(Q)$ is a *Müller acceptance condition.*

The *run* of the tree automaton $M$ on a $\Delta$-tree $t$ is the (uniquely determined) mapping $r_t : \mathrm{dom}(t) \to Q$ satisfying the following conditions:

- $r_t(\varepsilon) = q_0$,
- for each inner node $u$ of $\mathrm{dom}(t)$, $\delta_k(r_t(u), t(u)) = \big(r_t(u \cdot 1), \ldots, r_t(u \cdot k)\big)$, where $k$ is the *rank* of $u$ in $t$, i.e., the unique integer with $t(u) \in \Delta_k$.

The automaton $M$ *accepts* the $\Delta$-tree $t$ if for each leaf $x \in \mathrm{dom}(t)$ of $t$, the pair $(r_t(x), t(x))$ is in $F_{\mathrm{fin}}$, moreover, for each infinite path $\pi = x_0, x_1 \ldots$ of $t$, the set

$$\{q \in Q : q \text{ occurs infinitely often in the sequence } (r_t(x_i))_i\}$$

is a member of $\mathcal{F}$. The *language accepted by M* is the set $L(M)$ of $\Delta$-trees accepted by $M$.

The following theorem[2] can be found in [13], Theorem 4.5:

**Theorem 7.** *The following decision problem is complete for* PSPACE*: given a Müller tree automaton M, does it hold that* $L(M) = \emptyset$*?*

Using this result, it can be shown that the emptiness problem of MCFLs (given by MCFGs $G = (V, \Sigma, P, S, \mathcal{F})$) is also PSPACE-complete. We define the *size* of an MCFG $G = (V, \Sigma, P, S, \mathcal{F})$ as the sum $|V| + |\Sigma| + \mathrm{size}(P) + \mathrm{size}(\mathcal{F})$, where $\mathrm{size}(P) = \sum_{A\to\alpha\in P}(|\alpha| + 1)$ and $\mathrm{size}(\mathcal{F}) = \sum_{F\in\mathcal{F}} |F|$.

Let $\Delta^{(G)}$ be the ranked alphabet defined as follows. Let $\Delta_0^{(G)}$ be the set $\Sigma \cup \{\bullet\}$ where $\bullet$ is a fresh symbol. Let $\Delta_1^{(G)}$ be the set $\{A \to \alpha \in P : |\alpha| \le 1\}$ of those rules of $G$ having right side of length *at most* 1. For each $k > 1$, let $\Delta_k^{(G)}$ be the set $\{A \to \alpha \in P : |\alpha| = k\}$ of those rules of $G$ having right side of length *exactly* $k$. Now let $\mathcal{R}$ be the set of those integers $k$ with $\Delta_k^{(G)}$ being nonempty and let $\Delta^{(G)}$ be the ranked alphabet $\biguplus_{k\in\mathcal{R}} \Delta_k^{(G)}$.

We define the Müller tree automaton $M_G = (Q, \Delta^{(G)}, \delta, q_0, F_{\mathrm{fin}}, \mathcal{F}')$, where

- $Q = V \cup \Sigma \cup \{\bullet, \perp\}$ and $\perp$ is a new state;
- $q_0 = S$, the start symbol;
- $F_{\mathrm{fin}} = \{(a, a) : a \in \Delta_0\}$;
- $\mathcal{F}' = \mathcal{F}$

and $\delta$ is given as follows: for each state $q \in Q$, integer $k \in \mathcal{R}$ and symbol $a \in \Delta_k^{(G)}$, let

$$\delta_k(q, a) = \begin{cases} (X_1, \ldots, X_k) & \text{if } q = A \in V,\ k > 0, \text{ and } a = A \to X_1 \ldots X_k \text{ for the symbols } X_1, \ldots, X_k \in V \cup \Sigma; \\ (\bullet) & \text{if } q = A \in V,\ k = 1 \text{ and } a = A \to \varepsilon; \\ (\perp, \ldots, \perp) & \text{otherwise.} \end{cases}$$

It is clear by the construction that a $\Delta^{(G)}$-tree $t$ is accepted by $M_G$ if and only if its $\Sigma \cup V \cup \{\varepsilon\}$-relabeling $t'$ defined by

$$t'(x) = \begin{cases} A & \text{if } t(x) = A \to \alpha \text{ for some } A \in V, \alpha \in (\Sigma \cup V)^*, \\ a & \text{if } t(x) = a \in \Sigma, \\ \varepsilon & \text{otherwise, i.e., when } t(x) = \bullet \end{cases}$$

is a complete derivation tree of $G$ with root symbol $S$ satisfying the Müller condition $\mathcal{F}$. Thus, $L^\infty(G)$ is empty if and only if $L(M_G)$ is empty, hence the emptiness problem of MCFLs is (logspace) reducible to the emptiness problem of languages accepted by (deterministic) Müller tree automata. Hence, the emptiness problem of MCFLs is in PSPACE.

Via a similar technique one can construct (in logspace) for a (nondeterministic) Müller tree automaton $M$ an MCFG $G_M$ such that $L(M)$ is empty if and only if $L^\infty(G_M)$ is empty, yielding that the emptiness problem of MCFLs is PSPACE-complete.

**Theorem 8.** *The following problem is complete for* PSPACE*: given an* MCFG *G, does it hold that* $L^\infty(G) = \emptyset$*?*

**Remark 9.** The problem is solvable in polynomial time for BCFGs; cf. [10].

## 5. A normal form

In this section we introduce a notion of normal form for MCFGs and show that each MCFG can be transformed into a normal form grammar in polynomial space. Given an MCFG $G = (V, \Sigma, P, S, \mathcal{F})$, we say that a nonterminal $A \in V$ is

- *accessible from a nonterminal B* if $B \Rightarrow^\infty uAv$ for some words $u, v \in \Sigma^\infty$; it is called *accessible* if it is accessible from $S$;
- *productive* if $A \Rightarrow^\infty u$ for some $u \in \Sigma^\infty$;
- $+$*-productive* if $A \Rightarrow^\infty u$ for some $u \in \Sigma^{+\infty}$;
- *useful* if it is either productive and accessible; or $A = S$ and $P \subseteq \{S \to \varepsilon\}$;
- *useless* if it is not useful.

Observe that $L^\infty(G)$ is empty for an MCFG $G = (V, \Sigma, P, S, \mathcal{F})$ with $P \not\subseteq \{S \to \varepsilon\}$ if and only if $S$ is useless. Thus, it is PSPACE-hard to decide whether a nonterminal $A$ of an MCFG $G$ is useless or not.

[2] Technically, the results of [13] concentrate only on *full binary trees*, i.e., when $\mathcal{R} = \{2\}$, but a standard logspace reduction exists from our more general notion of trees to the case of full binary trees.

An MCFG $G = (V, \Sigma, P, S, \mathcal{F})$ is in *normal form* if it satisfies the following conditions:

- Either $V = \{S\}$ and $P \subseteq \{S \to \varepsilon\}$ (in which case $L^\infty(G)$ is either $\emptyset$ or $\{\varepsilon\}$), or $G$ only contains nonterminals which are both $+$-productive and accessible (in which case $L^\infty(G)$ contains a nonempty word).
- Whenever $A \in V$ is a nonterminal with $A \Rightarrow^\infty \varepsilon$, then $A \to \varepsilon$ is a production in $P$.

We call a derivation tree $t$ of $G$ *locally finite* if for each node $x \in \mathrm{dom}(t)$, if $t|_x$ is infinite then $\mathrm{fr}(t|_x) \neq \varepsilon$.

**Proposition 10.** *Suppose $G = (V, \Sigma, P, S, \mathcal{F})$ is an* MCFG *in normal form and $t$ is a derivation tree of $G$. Then there exists a locally finite derivation tree $t'$ of $G$ with the same root symbol and the same frontier.*

**Proof.** Let $t$ be a derivation tree of $G$. Let $X \subseteq \mathrm{dom}(t)$ denote the set of those nonterminal labeled nodes of $t$ whose corresponding subtree has empty frontier, i.e., $X = \{x \in \mathrm{dom}(t) : t(x) \in V, \mathrm{fr}(t|_x) = \varepsilon\}$. Let $X_0 \subseteq X$ be the set of all minimal nodes of $X$ with respect to the prefix ordering. Then,

$$X_0 = \{x \in \mathrm{dom}(t) : t(x) \in V, \mathrm{fr}(t|_x) = \varepsilon, \mathrm{fr}(t|_y) \neq \varepsilon \text{ for all } y <_{\mathrm{pr}} x\}.$$

Let $t'$ denote the tree obtained from $t$ by substituting the tree $t_x$ for each node $x \in X_0$, having two nodes, a root labeled by $t(x)$ and a leaf labeled by $\varepsilon$. Then for every $x \in X_0$, since $t(x) \Rightarrow^\infty \varepsilon$ and since $G$ is in normal form, $t(x) \to \varepsilon$ is a production. Thus $t'$ is a derivation tree of $G$. Moreover, $t'$ is a locally finite derivation tree having the same frontier and root symbol as $t$, proving the claim. □

**Proposition 11.** *There exists a polynomial-space algorithm which constructs for an arbitrary* MCFG *$G$ an equivalent* MCFG *$G_u$ containing no useless nonterminals, moreover, the size of the resulting grammar $G_u$ is at most the size of the original grammar $G$.*

**Proof.** Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an MCFG. By definition, a nonterminal $A \in V$ is productive if and only if $L^\infty(G, A)$ is nonempty. Since by Theorem 8 the emptiness problem of MCFLs given by an MCFG is decidable in polynomial space, the set $V_p \subseteq V$ of productive nonterminals can be computed using polynomial space. If $S \notin V_p$, then $L^\infty(G)$ is empty and $G_u = (\{S\}, \Sigma, \emptyset, S, \emptyset)$ is an MCFG equivalent to $G$ containing no useless nonterminals.

Otherwise, if $S \in V_p$, it is easy to see that a nonterminal $A$ is useful if and only if $A \in V_p$ and $S \Rightarrow^* \alpha A \beta$ for some $\alpha, \beta \in (V_p \cup \Sigma)^*$. Given $V_p$, this condition can be decided in polynomial time, by the usual testing of reachability, thus the set $V_u$ of useful nonterminals can be computed in polynomial space.

The grammar $G_u = (V_u, \Sigma, P_u, S, \mathcal{F}_u)$ is equivalent to $G$, where $P_u \subseteq P$ consists of those rules $A \to \alpha \in P$ with $A \in V_u$ and $\alpha \in (V_u \cup \Sigma)^*$, and $\mathcal{F}_u \subseteq \mathcal{F}$ consists of those sets $F \in \mathcal{F}$ which are subsets of $V_u$. Moreover, $G_u$ does not contain useless nonterminals. □

Observe that for an MCFG $G$ containing no useless nonterminals we have that $L^\infty(G) = \emptyset$ if and only if its set of productions is empty. Since the emptiness problem of the generated language is PSPACE-complete for arbitrary MCFGs, the above polynomial-space algorithm is the best one we can hope for, under standard assumptions of complexity theory.

**Proposition 12.** *There exists a polynomial-time algorithm which constructs for an* MCFG *$G$ containing no useless nonterminals an equivalent* MCFG *$G'$ in normal form. Moreover, the size of the resulting grammar $G'$ is at most the size of $G$.*

**Proof.** Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an MCFG containing no useless nonterminals. If $P \subseteq \{S \to \varepsilon\}$, then $G$ is already in normal form and we may choose $G' = G$. Otherwise, we compute the set $V_0 \subseteq V$ of those nonterminals $A$ for which $L^\infty(G, A) = \{\varepsilon\}$ (i.e. the set of those nonterminals which are *not* $+$-productive) using the fact that $A \notin V_0$ if and only if $A \Rightarrow^* \alpha$ for some $\alpha \in (\Sigma \cup V)^*$ containing at least one terminal symbol (here we use that $G$ contains only productive nonterminals). This condition can be decided in polynomial time for each $A \in V$, hence the set $V_0$ can also be computed in polynomial time.

If $S \in V_0$, then the grammar $(\{S\}, \Sigma, \{S \to \varepsilon\}, S, \emptyset)$ is an MCFG in normal form which is equivalent to $G$. Otherwise, we construct the set $V_\varepsilon \subseteq V$ of all nonterminals $A$ with $A \Rightarrow^\infty \varepsilon$ (the set of *nullable* nonterminals). To this end, let $P_\varepsilon \subseteq P$ be the set of rules $A \to \alpha \in P$ with $\alpha \in V^*$, and consider the grammar $G_{A,\varepsilon} = (V, \Sigma, P_\varepsilon, A, \mathcal{F})$ for each nonterminal $A \in V$.

We claim that $L^\infty(G_{A,\varepsilon})$ is nonempty if and only if $A$ is nullable. Indeed,

$A$ is nullable $\Leftrightarrow$ $\mathrm{fr}(t) = \varepsilon$ for some derivation tree $t$ of $G$ with root symbol $A$
$\Leftrightarrow$ there exists some complete derivation tree $t$ of $G$ with root symbol $A$ in which no terminal occurs as (leaf) symbol
$\Leftrightarrow$ there exists some complete derivation tree $t$ of $G_{A,\varepsilon}$ with root symbol $A$
$\Leftrightarrow$ $L^\infty(G_{A,\varepsilon})$ is nonempty.

Hence, the set $V_\varepsilon$ can be computed using polynomial space.

Then, the MCFG $(V', \Sigma, P', S, \mathcal{F}')$ where

- $V' = V - V_0$,
- $A \to X_1 \dots X_n$ is a production in $P'$ for some $n > 0$ and $X_1, \dots, X_n \in (V - V_0) \cup \Sigma$ if and only if $A \to \alpha_0 X_1 \alpha_1 \dots X_{n-1} \alpha_{n-1} X_n \alpha_n$ is a production in $P$ for some $\alpha_0, \dots, \alpha_n \in V_0^*$,

- $A \to \varepsilon$ is a production in $P'$ if and only if $A \in V_\varepsilon$,
- $\mathcal{F}'$ consist of those sets $F \in \mathcal{F}$ which are subsets of $V - V_0$,

is an MCFG equivalent to $G$ which is in normal form. $\square$

Thus we can formulate the main result of this section:

**Theorem 13.** *Given an arbitrary* MCFG *$G$, one can construct an equivalent* MCFG *$G'$ in normal form using polynomial working space, moreover, the size of the resulting $G'$ is at most the size of $G$.*

## 6. Decision problems

In this section we investigate the complexity of the following decision problems: given an MCFG $G$ in normal form, does it hold for the language $L = L^\infty(G)$ that

- $L$ is empty;
- $L$ contains at least one infinite word;
- $L$ contains only well-ordered words;
- $L$ contains only finite and $\omega$-words;
- $L$ contains only scattered words;
- $L$ contains only dense words.

Each of these problems can be shown to be decidable by using the fact that languages recognized by Müller tree automata are effectively closed under certain operations such as the Boolean operations and the emptiness problem is decidable for Müller tree automata. However, this general method does not provide effective algorithms and the simple characterizations obtained by the direct methods developed in this section. The main results of this section give a nontrivial upper bound for the complexity of the above questions: all these questions are solvable in polynomial time if the grammar $G$ is in normal form. By Theorem 13 this yields a polynomial-space upper bound for these questions in the general case.

The first result of this section is the complexity of the emptiness problem, which is almost trivial when the grammar is in normal form.

**Theorem 14.** *The following problem can be decided in polynomial time: given an* MCFG $G = (V, \Sigma, P, S, \mathcal{F})$ *with no useless nonterminals, is $L^\infty(G)$ empty?*

**Proof.** As we already argued, $L^\infty(G) = \emptyset$ if and only if $P = \emptyset$, which is clearly decidable in polynomial time. $\square$

In this section we will use the following lemma several times.

**Lemma 15.** *Suppose $t$ is a locally finite derivation tree of an* MCFG *$G$. Then $t$ is infinite if and only if* $\mathrm{fr}(t)$ *is an infinite word.*

When $G = (V, \Sigma, P, S, \mathcal{F})$ is an MCFG, we let $L^*(G)$ denote the set $\{w \in \Sigma^* : S \Rightarrow^* w\}$. (Equivalently, $L^*(G)$ could be defined as $L^\infty(G')$ for $G' = (V, \Sigma, P, S, \emptyset)$.)

**Proposition 16.** *If $G = (V, \Sigma, P, S, \mathcal{F})$ is an* MCFG *in normal form, then $L^*(G) = L^\infty(G) \cap \Sigma^*$. Hence, it can be decided in polynomial time whether an* MCFL *given by an* MCFG *in normal form contains at least one finite word.*

**Proof.** Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an MCFG in normal form. It is clear that $L^*(G) \subseteq L^\infty(G) \cap \Sigma^*$. To show the opposite inclusion, let $w \in L^\infty(G)$ be a finite word. Since $G$ is in normal form, by Proposition 10, $w = \mathrm{fr}(t)$ for some locally finite derivation tree $t$ of $G$ with root symbol $S$. Since $t$ is a locally finite derivation tree having a finite frontier, by Lemma 15 $t$ is finite as well, hence $w \in L^*(G)$, proving the claim. $\square$

**Corollary 17.** *A language $L \subseteq \Sigma^*$ of finite words is context-free if and only if it is an* MCFL.

Next we describe a construction that will be often used in the sequel. Given an MCFG $G = (V, \Sigma, P, S, \mathcal{F})$, we define a finite edge-labeled multigraph $\Gamma_G$ as follows:

- The set of nodes of $\Gamma_G$ is $V \cup \Sigma$.
- The edge labels are pairs of words of the form $(\alpha, \beta)$ with $\alpha, \beta \in (V \cup \Sigma)^*$.
- There exists an edge from $A \in V \cup \Sigma$ to $B \in V \cup \Sigma$ labeled $(\alpha, \beta)$, in notation $A \xrightarrow{\alpha,\beta} B$ if and only if $A \to \alpha B \beta$ is a production in $P$ (hence each terminal node of $\Gamma_G$ is a sink).

For an arbitrary set $X \subseteq V \cup \Sigma$ of symbols let $\Gamma_G|_X$ denote the restriction of $\Gamma_G$ onto the set $X$, i.e., for the edge-labeled multigraph with node set $X$ and edges $A \xrightarrow{\alpha,\beta} B$ if and only if $A, B \in X$ and $A \to \alpha B \beta \in P$. The graph $\Gamma_G$ is a useful auxiliary

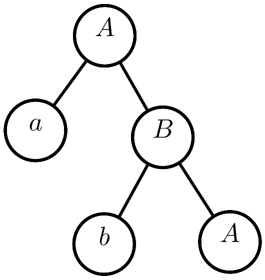


**Fig. 1.** A derivation tree $t$.

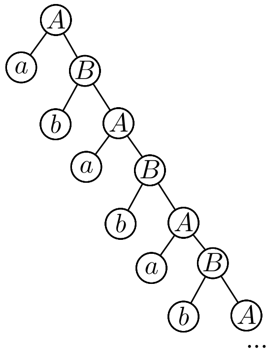


**Fig. 2.** The tree $(t, x)^\omega$ with $x = 22$, the lower right leaf of $t$.

construction as shown below. First, observe that a symbol $A$ is accessible from a symbol $B$ if and only if $A$ is accessible from $B$ in the graph $\Gamma_G$.

Let us call a set $F \in \mathcal{F}$ of nonterminals *viable* if there exists a derivation tree $t$ of $G$ with root symbol $S$ and an infinite path $\pi$ of $t$ such that $\mathrm{InfLab}(\pi) = F$.

When $t$ is a derivation tree of $G$ and $\varepsilon \neq x \in \mathrm{dom}(t)$ is a leaf node of $t$ with $t(\varepsilon) = t(x)$, we define the tree $(t, x)^\omega$ as follows: let $\mathrm{dom}((t, x)^\omega) = \{x^n \cdot y : n \geq 0, y \in \mathrm{dom}(t)\}$ and $(t, x)^\omega(x^n \cdot y) = t(y)$ for each $n \geq 0$ and $y \in \mathrm{dom}(t)$.

As an example, consider the following MCFG:

$$\begin{aligned} G &= (\{S, A, B\}, \{a, b\}, P, S, \mathcal{F}) \text{ where} \\ P &= \{S \to \varepsilon, S \to A, A \to aB, A \to SB, B \to bA\} \text{ and} \\ \mathcal{F} &= \{\{A, B\}, \{A\}\}. \end{aligned}$$

Fig. 1 shows a derivation tree $t$ of $G$ having leaf node $x = 22$ with $t(\varepsilon) = t(22) = A$. Observe that $\{t(\varepsilon), t(2), t(22)\} = \{A, B\}$ is an accepting set. Fig. 2 shows the tree $(t, x)^\omega$ which is also a derivation tree of $G$, showing $\{A, B\}$ is a viable set. On the other hand, $\{A\} \in \mathcal{F}$ is *not* a viable set of $G$.

It is clear that $(t, x)^\omega$ is a derivation tree of $G$ if and only if $F = \{t(x') : x' \leq_{\mathrm{pr}} x\}$ is in $\mathcal{F}$, in which case $\mathrm{InfLab}(\pi) = F$ for the infinite path $\pi$ of $(t, x)^\omega$ with $\mathrm{dom}(\pi) = \{x^n \cdot x' : n \geq 0, x' \leq_{\mathrm{pr}} x\}$. Hence, a set $F \in \mathcal{F}$ is viable if (and only if) $F = \{t(x') : x' \leq_{\mathrm{pr}} x\}$ for some derivation tree $t$ of $G$ and leaf node $x$ of $t$ such that $t(x) = t(\varepsilon)$ is an accessible nonterminal of $G$.

The following fact can be easily proved by induction on $k$.

**Lemma 18.** *Suppose $A_0 \xrightarrow{\alpha_1, \beta_1} A_1 \xrightarrow{\alpha_2, \beta_2} \dots \xrightarrow{\alpha_k, \beta_k} A_k$ is a path in $\Gamma_G$. Then there exists a (finite) derivation tree $t$ of $G$ and a leaf node $x$ of $t$ with*

$$\mathrm{fr}(t) = \alpha_1 \alpha_2 \dots \alpha_k A_k \beta_k \dots \beta_2 \beta_1,$$

*$t(\varepsilon) = A_0$, $t(x) = A_k$ and $\{t(x') : x' \leq_{\mathrm{pr}} x\} = \{A_0, \dots, A_k\}$.*

**Proposition 19.** *Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an* MCFG *without useless nonterminals. A set $F \in \mathcal{F}$ is viable in $G$ if and only if $\Gamma_G|_F$ is strongly connected. Hence, viability of a set $F \in \mathcal{F}$ can be decided in polynomial time.*

**Proof.** Let $F \in \mathcal{F}$ be a viable set. Then $F = \mathrm{InfLab}(\pi)$ for some infinite path $\pi = x_0, x_1, \dots$ of a derivation tree $t$ of $G$ with root symbol $S$. Clearly, there exist integers $0 \leq i < j$ such that $t(x_i) = t(x_j)$ and $\{t(x_i), \dots, t(x_j)\} = F$. By definition of $\Gamma_G$, whenever a node $y$ is a successor of a node $x$ in a derivation tree $t'$ of $G$, and $t'(y)$ (hence also $t'(x)$) is a nonterminal symbol, then $t'(x) \xrightarrow{\alpha, \beta} t'(y)$ is an edge in the graph $\Gamma_G$ for some words $\alpha, \beta$. Hence there exists a closed path $t(x_i) \xrightarrow{\alpha_i, \beta_i} t(x_{i+1}) \xrightarrow{\alpha_{i+1}, \beta_{i+1}} \dots \xrightarrow{\alpha_{j-1}, \beta_{j-1}} t(x_j)$, showing $\Gamma_G|_F$ is strongly connected.

For the other direction, suppose $\Gamma_G|_F$ is strongly connected. Let

$$A_0 \xrightarrow{\alpha_1,\beta_1} A_1 \xrightarrow{\alpha_2,\beta_2} \dots \xrightarrow{\alpha_k,\beta_k} A_k$$

be a closed path in $\Gamma_G|_F$ visiting each node of $F$ at least once, i.e., $\{A_0, \dots, A_k\} = F$ and $A_0 = A_k$. Applying Lemma 18 we get that there exists a derivation tree $t$ of $G$ with root symbol $A_0$ and a leaf node $x$ with $t(x) = A_0$ such that $F = \{t(x') : x' \le_{\mathrm{pr}} x\}$. Thus, for the infinite path $\pi = \{x^n \cdot x' : n \ge 0, x' \le_{\mathrm{pr}} x\}$ of $(t, x)^\omega$ we have $\mathrm{InfLab}(\pi) = F$. Since $G$ has no useless nonterminals, it follows now that $F$ is viable. □

**Theorem 20.** *Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an* MCFG *in normal form and let $X \in V$ be a nonterminal. $L^\infty(G, X)$ contains an infinite word if and only if there exists a viable set $F \in \mathcal{F}$ such that some (hence each) member of $F$ is accessible from $X$ and there is an edge $A \xrightarrow{\alpha,\beta} B$ in $\Gamma_G$ with $A, B \in F$ and $\alpha\beta \ne \varepsilon$. Hence, it can be decided in polynomial time whether an* MCFL *given by an* MCFG *in normal form contains finite words only.*

**Proof.** Suppose $L^\infty(G, X)$ contains an infinite word $w$ and let $t$ be a derivation tree of $G$ with $\mathrm{fr}(t) = w$ and $t(\varepsilon) = X$. Let $\pi = x_0, x_1, \dots$ be an infinite path of $t$ such that $\mathrm{fr}(t|_{x_i})$ is an infinite word for each $i \ge 0$ (such a path always exists). Let $F = \mathrm{InfLab}(\pi)$: then $F \in \mathcal{F}$ is a viable set and its members are obviously accessible from $X$. Let $i > 0$ be an integer with $\{t(x_k) : k \ge i\} = F$. Since $\mathrm{fr}(t|_{x_i}) \ne \varepsilon$, there exists a descendant $x_j \in \pi$ of $x_i$ with $\deg_t(x_j) > 1$. Let $A = t(x_j)$ and $B = t(x_{j+1})$, so that both $A$ and $B$ are in $F$. Since $t$ is a derivation tree of $G$, there is a production $A \to \alpha B\beta$ in $P$ for some $\alpha, \beta \in (V \cup \Sigma)^*$ with $\alpha\beta \ne \varepsilon$. Hence, $A \xrightarrow{\alpha,\beta} B$ is an edge of $\Gamma_G$ satisfying the claim.

For the other direction, suppose $A \xrightarrow{\alpha,\beta} B$ is an edge of $\Gamma_G$ with $A, B \in F$ for some viable set $F \in \mathcal{F}$ and $\alpha\beta \ne \varepsilon$. Then $\Gamma_G|_F$ is strongly connected, thus there exists a path $A_0 \xrightarrow{\alpha_1,\beta_1} A_1 \xrightarrow{\alpha_2,\beta_2} \dots \xrightarrow{\alpha_k,\beta_k} A_k$ with $A_0 = B, A_k = A$ and $\{A_1, \dots, A_k\} = F$. Hence, there exists a derivation tree $t$ and a leaf $x$ of $t$ with $\mathrm{fr}(t) = \alpha_1 \dots \alpha_k \alpha B \beta \beta_k \dots \beta_1$, $t(\varepsilon) = t(x) = B$ and $\{t(x') : x' \le x\} = F$. Then the tree $t' = (t, x)^\omega$ is also a derivation tree of $G$ with $t'(\varepsilon) = B$ and $\mathrm{fr}(t') = (\alpha_1 \dots \alpha_k \alpha)^\omega (\beta\beta_k \dots \beta_1)^{-\omega}$. Since $\alpha\beta$ is nonempty and each nonterminal of $G$ is $+$-productive, there exist words $u, v \in \Sigma^\infty$ with $uv \ne \varepsilon$ such that $\alpha_1 \dots \alpha_k \alpha \Rightarrow^\infty u$ and $\beta\beta_k \dots \beta_1 \Rightarrow^\infty v$, thus $B \Rightarrow^\infty u^\omega v^{-\omega}$, which is an infinite word since $uv$ is nonempty. Since $B$ is accessible from $X$, there exist words $u_0, v_0$ with $X \Rightarrow^\infty u_0 B v_0$, showing $u_0 u^\omega v^{-\omega} v_0$ is an infinite word in $L^\infty(G, X)$. □

**Theorem 21.** *Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an* MCFG *in normal form. $L^\infty(G)$ contains a word which is not well-ordered if and only if there exists a viable set $F \in \mathcal{F}$ and an edge $A \xrightarrow{\alpha,\beta} B$ in $\Gamma_G|_F$ with $\beta \ne \varepsilon$. Hence, it can be decided in polynomial time whether an* MCFL *given by an* MCFG *in normal form contains well-ordered words only.*

**Proof.** Suppose $w \in L^\infty(G)$ is a word which is not well-ordered. Let $t$ be a derivation tree of $G$ with $\mathrm{fr}(t) = w$ and $t(\varepsilon) = S$. Then there exists an infinite sequence $x_0 >_{\mathrm{lex}} x_1 >_{\mathrm{lex}} \dots$ of leaves of $t$ such that $t(x_i)$ is a terminal symbol for each $i \ge 0$.

Let us construct an infinite path $\pi = y_0, y_1, \dots$ of $t$ as follows: let $y_0 = \varepsilon$. Suppose we have already constructed $y_i$ in such a way that for some $k \ge 0$, every $x_{k'}$ with $k' \ge k$ is a descendant of $y_i$. (This holds for $i = 0$ with $k = 0$.) Then let $y_{i+1} = y_i \cdot j$ where $j$ is the least integer with $y_i \cdot j$ being an ancestor of some $x_{k_1}$. Such a successor always exists and it is clear that for any $k' \ge k_1$, $x_{k'}$ is a descendant of $y_{i+1}$.

Let $F$ be the set $\mathrm{InfLab}(\pi)$ of nonterminals. Since $t$ is a derivation tree of $G$, $F \in \mathcal{F}$ is a viable set. For each $i$, let $y_{j_i} <_{\mathrm{pr}} x_i$ be the last ancestor of $x_i$ lying on $\pi$. Then there exists an integer $k$ such that the set $\{t(y_{k'}) : k' \ge j_k\}$ is $F$. By construction, $A = t(y_{j_k})$ and $B = t(y_{j_k+1})$ belong to $F$, and since $y_{j_k} <_{\mathrm{pr}} x_k$ but $y_{j_k+1} \not<_{\mathrm{pr}} x_k$, there exists an edge $A \xrightarrow{\alpha,\beta} B$ in $\Gamma_G|_F$ for some $\alpha, \beta \in (V \cup \Sigma)^*$ with $\beta \ne \varepsilon$.

For the other direction, let $F \in \mathcal{F}$ be a viable set, $A, B \in F$ and $A \xrightarrow{\alpha,\beta} B$ be an edge of $\Gamma_G$ with $\beta \ne \varepsilon$. Since $\Gamma_G|_F$ is strongly connected, there exists a path $A_0 \xrightarrow{\alpha_1,\beta_1} A_1 \xrightarrow{\alpha_2,\beta_2} \dots \xrightarrow{\alpha_k,\beta_k} A_k$ in $\Gamma_G$ with $A_0 = B, A_k = A$ and $\{A_0, \dots, A_k\} = F$. Hence, there exists a derivation tree $t$ and a leaf $x$ of $t$ with $\mathrm{fr}(t) = \alpha_1 \dots \alpha_k \alpha B \beta \beta_k \dots \beta_1$, $t(\varepsilon) = t(x) = B$ and $\{t(x') : x' \le x\} = F$. Then the tree $t' = (t, x)^\omega$ is also a derivation tree of $G$ with $t'(\varepsilon) = B$ and $\mathrm{fr}(t') = (\alpha_1 \dots \alpha_k \alpha)^\omega (\beta\beta_k \dots \beta_1)^{-\omega}$. Since $\beta$ is nonempty and each nonterminal of $G$ is $+$-productive, there exist words $u, v \in \Sigma^\infty$ with $v \ne \varepsilon$ such that $\alpha_1 \dots \alpha_k \alpha \Rightarrow^\infty u$ and $\beta\beta_k \dots \beta_1 \Rightarrow^\infty v$, so that $B \Rightarrow^\infty u^\omega v^{-\omega}$. Since $B$ is also accessible, there exist words $u_0, v_0 \in \Sigma^\infty$ with $S \Rightarrow^\infty u_0 B v_0$, thus $u_0 u^\omega v^{-\omega} v_0$ is a word in $L^\infty(G)$ which is not well-ordered. □

**Theorem 22.** *Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an* MCFG *in normal form. $L^\infty(G)$ contains an infinite word which is not an $\omega$-word if and only if there exists a viable set $F \in \mathcal{F}$, nonterminals $A, B \in V$ satisfying the following conditions:*

- *there exists an edge $A \xrightarrow{\alpha,\beta} B$ with $\beta \ne \varepsilon$;*
- *some (and hence each) member of $F$ is accessible from $B$ in $\Gamma_G$;*
- *there exists some edge $C \xrightarrow{\alpha',\beta'} D$ in $\Gamma_G|_F$ with $\alpha'\beta' \ne \varepsilon$.*

*Hence, it can be decided in polynomial time whether an* MCFL *given by an* MCFG *in normal form contains finite and $\omega$-words only.*

**Proof.** Note that an infinite word $w$ is an $\omega$-word if and only if for any factorization $w = u \cdot v$, if $u$ is infinite then $v$ is empty. Hence, $L^\infty(G)$ contains an infinite word which is not an $\omega$-word if and only if there exists a derivation tree $t$, a node $x \in \mathrm{dom}(t)$ and integers $i < j$ such that $x \cdot i$ and $x \cdot j$ are both in $\mathrm{dom}(t)$, $\mathrm{fr}(t|_{x\cdot i})$ is an infinite word and $\mathrm{fr}(t|_{x\cdot j})$ is nonempty.

So suppose $t$ is such a tree and let $A = t(x)$, $B = t(x \cdot i)$. Then, since $x \cdot j$ is also a successor of $x$ in $t$ and $t$ is a derivation tree of $G$, there exists an edge $A \xrightarrow{\alpha,\beta} B$ in $\Gamma_G$ with $\beta \neq \varepsilon$. Since $\mathrm{fr}(t|_{x\cdot i})$ is infinite, applying Theorem 20, it follows that there exists a viable set $F \in \mathcal{F}$ such that each member of $F$ is accessible from $B$, moreover, there exists an edge $C \xrightarrow{\alpha',\beta'} D$ in $\Gamma_G|_F$ with $\alpha'\beta' \neq \varepsilon$.

For the other direction, suppose $F \in \mathcal{F}$, $A, B \in V$, $C, D \in F$ and $\alpha, \beta, \alpha', \beta' \in (V \cup \Sigma)^*$ satisfy the conditions of the Theorem. Applying Theorem 20 we get that $C \Rightarrow^\infty w$ for some infinite word $w \in \Sigma^\infty$. Since each element of $F$ is accessible from $B$, for some word $w' = uwv$ with $u, v \in \Sigma^\infty$ it holds that $B \Rightarrow^\infty w'$. Hence, since $A \to \alpha B\beta$ is a production in $P$, and $\beta \neq \varepsilon$, for some word $w''$ of the form $w'' = u_1 w' u_2$ with $u_1, u_2 \in \Sigma^\infty$, $u_2 \neq \varepsilon$ we have $A \Rightarrow^\infty u_1 w' u_2$. Finally, since $A$ is accessible, $S \Rightarrow^\infty u_0 u_1 w' u_2 u_3$ holds for some $u_0, u_3 \in \Sigma^\infty$, which is infinite but not an $\omega$-word, since it can be written as $u_0 u_1 w' \cdot u_2 u_3$, where the first factor is infinite and the second is a nonempty word. $\square$

In the rest of the section our aim is to show that it is also decidable whether an MCFL given by an MCFG contains a quasi-dense word or consists of scattered words only, or whether it consists of dense words only. We also establish a property of those MCFLs which contain scattered words only.

**Theorem 23.** *Let $G = (V, \Sigma, P, S, \mathcal{F})$ be an* MCFG *in normal form. The following are equivalent:*

1. *$L^\infty(G)$ contains a quasi-dense word (i.e., a word which is not scattered).*
2. *There is a locally finite derivation tree $t$ (whose frontier is in $\Sigma^\infty$ and root symbol is $S$) such that the full (infinite) binary tree can be embedded in $t$.*
3. *There is a derivation tree $t$ (whose frontier is in $\Sigma^\infty$ and root symbol is $S$) such that the full (infinite) binary tree can be embedded in $t$.*
4. *There is a nonterminal $A$ and a finite derivation tree $t$ with root label $A$ which has two leaves $x_1$ and $x_2$ labeled $A$ with the following property: there is a set $F \in \mathcal{F}$ such that the set of labels of nonterminals along the path from the root to $x_i$ is equal to $F$, for $i = 1, 2$.*

*Hence, it is decidable in polynomial time whether $L^\infty(G)$ contains scattered words only.*

**Proof.** Suppose that $L^\infty(G)$ contains a quasi-dense word. Then there is a locally finite derivation tree $t$ with root label $S$ whose frontier is a quasi-dense word in $L^\infty(G)$. We define a set of nodes of $t$ which determines a full binary tree. To this end, let $X_0$ be the singleton set consisting only of the root. Suppose that $X_n$ has already been defined so that the nodes in $X_n$ determine a complete binary tree $t_n$ of depth $n$ and such that the frontier of the subtree rooted at each leaf of $t_n$ is a quasi-dense word. Consider a leaf $x$ of $t_n$. There exists a (not necessarily proper) descendant of $x$ which has two different successors $x_1$ and $x_2$ such that the frontier of the subtrees rooted at both successors are quasi-dense. Indeed, in the opposite case we could construct an infinite path $\pi$ of $t|_x$ such that the frontier of a subtree of $t|_x$ is quasi-dense if and only if its root belongs to $\pi$, a contradiction. (See also [3].) We construct $X_{n+1}$ by adding, for each leaf $x$ of $t_n$, two such nodes $x_1$ and $x_2$ to $X_n$. Finally let $X = \bigcup_{n\geq 0} X_n$. It is clear that the nodes in $X$ determine an embedding of the full binary tree in $t$. Thus, the first condition implies the second.

It is clear that the second condition implies the third. We now show that the third condition implies the first. To this end, assume that $t$ is a derivation tree which contains an embedded full binary tree. Since there are no useless nonterminals, without loss of generality we may assume that the frontier of $t$ is a terminal word and that the root of $t$ is labeled $S$. Now $t$ also contains a full ternary tree $t_0$ (since the full ternary tree can be embedded into the full binary tree) whose nodes form a subset $X$ of the node set of $t$ and whose root is $x_0$. We will modify the tree $t$ to obtain a derivation tree $t'$ whose frontier is a quasi-dense word in $\Sigma^\infty$. To this end, consider any non-root node $x \in X$ which is the second (middle) successor of a node in $X$ in the full ternary tree $t_0$. For each such node $x$ labeled $A_x$, we replace the subtree of $t'$ rooted $x$ with a derivation tree whose frontier is a *nonempty* word in $\Sigma^\infty$. It is clear that the resulting tree is a derivation tree whose root symbol is $S$ and whose frontier is a quasi-dense word in $\Sigma^\infty$.

We have thus proved that the first three conditions are equivalent. Suppose now that there is a locally finite derivation tree $t$ such that the full binary tree can be embedded in $t$. Since the set of nonterminals is finite, there is a set $X$ of nodes of $t$ that determines a full binary tree in $t$ such that each node in $X$ is labeled by the same nonterminal $A$. Indeed, let $t'$ be a copy of the full binary tree in $t$ whose set of node labels $V_0$ is minimal. Then for each nonterminal $A \in V_0$ and each node $x$ of $t'$, both successors of $x$ have descendants in $t'$ labeled $A$.

Since $t$ is locally finite, it follows as above that the frontier of each subtree rooted at a node in $X$ is quasi-dense. Now for each node $x \in X$, let $P_x$ denote the set of all infinite paths of $t|_x$ visiting the set $X$ infinitely often. Moreover, for each $x \in X$ consider the set $\mathcal{F}_x$ of all sets $F \subseteq V$ that arise as the set of labels of nodes along a path in $P_x$. Thus $\mathcal{F}_x$ is a nonempty subset of nonempty subsets of $V$.

Let us introduce the following quasi-order $\sqsubseteq$ on the nonempty subsets of $V$: Define $\mathcal{F}_1 \sqsubseteq \mathcal{F}_2$ if and only if for each $F_1 \in \mathcal{F}_1$ there is some $F_2 \in \mathcal{F}_2$ with $F_1 \subseteq F_2$. Moreover, define $\mathcal{F}_1 \equiv \mathcal{F}_2$ if and only if $\mathcal{F}_1 \sqsubseteq \mathcal{F}_2$ and $\mathcal{F}_2 \sqsubseteq \mathcal{F}_1$. The following holds:

*Claim.* For $\mathcal{F}_1$ and $\mathcal{F}_2$ as above, if $\mathcal{F}_1 \equiv \mathcal{F}_2$, then the set of maximal sets (with respect to set inclusion) in $F_1$ is equal to the set of maximal sets in $\mathcal{F}_2$, in notation, $\mathcal{F}_1^{\max} = \mathcal{F}_2^{\max}$.

It is clear that whenever $y \in X$ is a descendant of $x \in X$, then $\mathcal{F}_y \sqsubseteq \mathcal{F}_x$. Let $\mathcal{F}_0$ denote a set which is minimal with respect to the quasi-order $\sqsubseteq$ among the sets $\mathcal{F}_x$, $x \in X$, so that for each $x \in X$, if $\mathcal{F}_x \sqsubseteq \mathcal{F}_0$ then $\mathcal{F}_x \equiv \mathcal{F}_0$. Consider a node $x_0 \in X$ with $\mathcal{F}_{x_0} = \mathcal{F}_0$. Let $X_0$ denote the set of those descendants of $x_0$ (including $x_0$) that belong to $X$. Whenever $x \in X_0$ we have that $\mathcal{F}_x \equiv \mathcal{F}_0$.

We claim that $\mathcal{F}_0^{\max} \subseteq \mathcal{F}$. Indeed, suppose that $F \in \mathcal{F}_0^{\max}$. Then for each $x \in X_0$, also $F \in \mathcal{F}_x^{\max}$. Thus, for each $x \in X_0$ there is an infinite path $\pi$ in $P_x$ such that the set of labels of nodes along $\pi$ is $F$. But then there is a nonempty initial segment $\pi_0$ of $\pi$ such that each member of $F$ already appears as the label of nodes along $\pi_0$. We can now extend $\pi_0$ to a finite initial segment of $\pi$ whose end node is in $X_0$: the set of labels does not change (by the maximality of $F$). So we have seen that for each $x \in X_0$ there is a node $y \in X_0$ which is a proper descendant of $x$ such that $F$ is the set of nonterminals that occur as the label of a node on the path from $x$ to $y$. This implies that for each $x \in X_0$ there is a path in $P_x$ such that the set of nonterminals that occur infinitely often as the label of a node along the path is exactly $F$. Since the tree $t$ is a derivation tree, it follows that $F$ is in $\mathcal{F}$.

We also claim that $\mathcal{F}_0^{\max}$ is a singleton set. Suppose that $F_1, F_2 \in \mathcal{F}_0^{\max}$. As explained above, for each $x \in X_0$ there exists a proper descendant $y \in X_0$ of $x$ such that the set of node labels along the path from $x$ to $y$ is $F_1$. Also, since $y \in X_0$, there also exists a proper descendant $z \in X_0$ of $y$ such that the set of node labels along the path from $y$ to $z$ is $F_2$. Thus, the set of node labels along the path from $x$ to $z$ is $F_1 \cup F_2$. This yields that $F_1 \cup F_2 \in \mathcal{F}_x^{\max} \equiv \mathcal{F}_0^{\max}$ and thus (by maximality of the sets $F_1$ and $F_2$) $F_1 = F_2$ and $\mathcal{F}_0^{\max}$ is indeed a singleton set. Let $F \in \mathcal{F}$ denote the unique element of $\mathcal{F}_0^{\max}$.

Now consider any two proper descendants $y_1$ and $y_2$ of $x_0$ in the set $X_0$, none of which is a descendant of the other. For each $i = 1, 2$ there is a descendant $x_i \in X_0$ of $y_i$ such that the set of node labels along the path from $y_i$ to $x_i$ is $F$. By maximality, the same holds for the path from $x_0$ to $x_i$, for $i = 1, 2$. Consider now the finite derivation tree whose nodes are those lying on the paths from $x_0$ to $x_i$, $i = 1, 2$, together with the direct descendants of all these nodes other than $x_1$ or $x_2$. This tree satisfies the fourth condition, proving that the third condition of the theorem implies the fourth.

Thus, it remains to prove that the fourth condition implies the third. But suppose that there is a finite derivation tree $t$ as described in the fourth condition. Then attaching a copy of $t$ to the nodes $x_1$ and $x_2$ and continuing this way up to infinity we can construct a derivation tree containing an embedded full binary tree. Now we attach a derivation tree whose frontier is a terminal word to each leaf labeled by a nonterminal. The resulting tree with root symbol $A$ has a terminal word as its frontier and contains an embedded full binary tree. Since $G$ contains no useless nonterminals, it follows that there is also a derivation tree containing an embedded full binary tree whose root symbol is $S$ and whose frontier is a word in $\Sigma^\infty$. $\square$

## 7. The case of scattered words

In this section our aim is to establish the following property of MCFLs: if $L$ is an MCFL consisting of scattered words, then one of the following conditions holds:

(i) either there exists an integer $n$ such that the *rank* of each member of $L$ is at most $n$,
(ii) or $L$ contains words having rank at least $\alpha$ for any countable ordinal $\alpha$.

Here our notion of rank is closely related to that of the Hausdorff-rank; cf. [14,16]. Moreover, it can be decided in polynomial time whether (i) or (ii) holds, and if (i) holds, the least such integer $n$ can also be computed, still in polynomial time (if $L$ is given by an MCFG in normal form). In order to prove these results, we introduce several notions.

For each countable ordinal $\alpha$ we define a set $H_\alpha$ of nonzero countable order types as follows.

1. $H_0$ consists of the finite nonzero order types.
2. $H_\alpha$ is the smallest set of order types which is closed under binary sum and contains all the order types of the form $\sum_{i\in I} \tau_i$, where the nonempty linear order $I$ is either finite or has order type $\omega$ or $-\omega$, and each $\tau_i$ is in $H_{\beta_i}$ for some ordinal $\beta_i < \alpha$.

It is known, cf. [16], that a nonempty countable ordering is scattered if and only if its order type is contained in $H_\alpha$ for some (countable) ordinal $\alpha$. Given a nonempty countable scattered ordering $P$, let $H(P)$ stand for its *rank*, i.e. the smallest ordinal $\alpha$ for which the order type of $P$ belongs to $H_\alpha$. When $u$ is a scattered word, its rank $H(u)$ is defined as the rank of its underlying linear order $o(u)$.

For an MCFG $G = (V, \Sigma, P, S, \mathcal{F})$ and symbols $X, Y \in V \cup \Sigma$, let us write $X \leadsto^\infty Y$ if and only if $X \Rightarrow^\infty u$ for some $u \in (V \cup \Sigma)^\infty$ with $Y$ occurring infinitely many times in $u$. We call a nonterminal $A \in V$ *reproductive* if $A \leadsto^\infty A$.

Below we will make use of the following known facts, see e.g. [14].

- if $P$ is a subordering of the countable scattered ordering $Q$, then $H(P) \leq H(Q)$, and
- if $P$ and $Q$ are countable scattered linear orderings, then $H(P \times Q) = H(P) + H(Q)$.

**Lemma 24.** *Suppose $G = (V, \Sigma, P, S, \mathcal{F})$ is an* MCFG *containing no useless nonterminals such that $L^\infty(G)$ consists of scattered words only. Suppose $A \leadsto^\infty B$ for the symbols $A, B \in V \cup \Sigma$ (thus $A \in V$). Then for any countable ordinal $\alpha$, if for each ordinal $\beta < \alpha$ there exists a word $u_\beta$ with $B \Rightarrow^\infty u_\beta$ and $H(u_\beta) \geq \beta$, then $A \Rightarrow^\infty u'$ for some word $u'$ with $H(u') \geq \alpha$.*

**Proof.** Suppose $A \leadsto^\infty B$ and let $u \in (V \cup \Sigma)^\infty$ be a word with $A \Rightarrow^\infty u$ in which $B$ occurs infinitely many times, let $\alpha$ be a countable ordinal and suppose that for each $\beta < \alpha$ there exists a word $u_\beta$ with $B \Rightarrow^\infty u_\beta$ and $H(u_\beta) \geq \beta$.

If $\alpha = \alpha' + 1$ is a successor ordinal, then by $A \Rightarrow^\infty u$ we have $A \Rightarrow^\infty u'$ where $u'$ can be constructed from $u$ by substituting $u_{\alpha'}$ for each occurrence of $B$ in $u$. Then, by the above facts, $H(u') \geq H(u_{\alpha'}) + 1 \geq \alpha$.

If $\alpha$ is a limit ordinal, then $\alpha = \sup_{i<\omega} \beta_i$ for some countable ordinals $\beta_0 < \beta_1 < \cdots < \alpha$. By assumption, for each $i < \omega$ there exists a word $u_{\beta_i}$ with $B \Rightarrow^\infty u_{\beta_i}$ and $H(u_{\beta_i}) \geq \beta_i$. Let $x_1, x_2, \ldots \in \mathrm{dom}(u)$ be an enumeration of the $B$-labeled positions of the word $u$ and let $u'$ be the word obtained from $u$ by substituting $u_{\beta_i}$ for each $x_i$, $i < \omega$. Clearly, $A \Rightarrow^\infty u'$. Also, $H(u') \geq \beta_i$ for each $i < \omega$, thus $H(u') \geq \alpha$. $\square$

An immediate corollary of Lemma 24 is:

**Corollary 25.** *Suppose $G = (N, \Sigma, P, S, \mathcal{F})$ is an* MCFG *in normal form such that $L = L^\infty(G)$ contains only scattered words and some nonterminal $A \in V$ is reproductive. Then for each countable ordinal $\alpha$, $L$ contains a word $u_\alpha$ with $H(u_\alpha) \geq \alpha$.*

**Proof.** Let $u, v \in \Sigma^\infty$ be words with $S \Rightarrow^\infty uAv$ (since the reproductive nonterminal $A$ is useful, such words exist). We show that for each countable ordinal $\alpha$ there exists a word $w_\alpha$ with $A \Rightarrow^\infty w_\alpha$ and $H(w_\alpha) \geq \alpha$, then choosing $u_\alpha = uw_\alpha v$ suffices.

We use transfinite induction. For $\alpha = 0$ the statement is clear, since by $+$-productivity of $A$ there exists a word $w_0 \in \Sigma^{+\infty}$ with $A \Rightarrow^\infty w_0$ and clearly, $H(w_0) \geq 0$ for any scattered nonempty word $w_0$.

Suppose the claim holds for all ordinals less than $\alpha$, i.e. for each $\beta < \alpha$ there exists a word $w_\beta$ with $A \Rightarrow^\infty w_\beta$ and $H(w_\beta) \geq \beta$. Applying Lemma 24 we get that the claim holds for $\alpha$, proving the statement. $\square$

The (transitive) relation $\rightsquigarrow^\infty$ is computable in polynomial time.

**Proposition 26.** *Suppose $G = (V, \Sigma, P, S, \mathcal{F})$ is an* MCFG *in normal form. Then $A \rightsquigarrow^\infty B$ holds for the symbols $A$ and $B$ if and only if there exists a viable set $F \in \mathcal{F}$ such that some (and hence each) member of $F$ is accessible from $A$, moreover, $B$ is accessible from some symbol occurring in at least one of the edge labels of $\Gamma_G|_F$.*

**Proof.** Suppose $A \rightsquigarrow^\infty B$ and let $t$ be a derivation tree of $G$ with $t(\varepsilon) = A$ and $\mathrm{fr}(t) = u$ for some word $u \in (V \cup \Sigma)^\infty$ such that $B$ occurs in $u$ infinitely many times. Then there exists an infinite path $\pi = x_0, x_1, \ldots$ of $t$ such that for each $i \geq 0$, $B$ occurs in $\mathrm{fr}(t|_{x_i})$ infinitely many times. Let $F \in \mathcal{F}$ be the set $\mathrm{InfLab}(\pi)$ of nonterminals. Clearly, $F$ is a viable set. By the constraint on $\pi$, there exists a node $x \in \pi$ such that $\{t(x') : x' \geq_{\mathrm{pr}} x, x' \in \pi\} = F$ and $B$ occurs in $\mathrm{fr}(t|_{x \cdot i})$ for some $i \in \mathbb{N}$ with $x \cdot i \in \mathrm{dom}(t) - \pi$. Let $j \neq i$ be the uniquely determined integer with $x \cdot j \in \pi$. Then there exists an edge $t(x) \xrightarrow{\alpha,\beta} t(x \cdot j)$ with $t(x)$, $t(x \cdot j)$ both in $F$ and $C = t(x \cdot i)$ occurring in $\alpha\beta$; since $B$ occurs in $\mathrm{fr}(t|_{x \cdot i})$, $B$ is also accessible from $C$.

For the other direction, suppose $F \in \mathcal{F}$ is a viable set of nonterminals, $C \xrightarrow{\alpha,\beta} D$ is an edge in $\Gamma_G|_F$, $X$ is a symbol occurring in $\alpha\beta$, $B$ is accessible from $X$ and some member of $F$ is accessible from $A$ (thus $C$ is also accessible from $A$).

Using similar arguments as before we get that there exists a derivation tree $t$ with root symbol $C$ and frontier word of the form $\alpha'^\omega \beta'^{-\omega}$ such that $\alpha'\beta' \in (V \cup \Sigma)^\infty$ contains an occurrence of $X$, thus $X$ occurs in $\mathrm{fr}(t)$ infinitely many times; since $B$ is accessible from $X$, we get $C \Rightarrow^\infty u$ for some word $u \in (V \cup \Sigma)^\infty$ in which $B$ occurs infinitely many times; finally, since $C$ is accessible from $A$, $A \Rightarrow^\infty u_0 u v_0$ for some words $u_0, v_0 \in (V \cup \Sigma)^*$, showing $A \rightsquigarrow^\infty B$. $\square$

The following is well-known.

**Lemma 27.** *The following are equivalent for a tree domain $T$:*

(i) *$T$ has only finitely many maximal paths;*
(ii) *whenever $T' \subseteq \mathrm{dom}(t)$ is a set of pairwise incomparable nodes of $T$ (with respect to the prefix ordering), then $T'$ is finite.*

**Proof.** It is clear that whenever $T' \subseteq T$ is a set of pairwise incomparable nodes of $T$, then for each $x \in T'$ there exists a maximal path $\pi_x$ containing $x$, moreover, since the elements of $T'$ are pairwise incomparable, these paths are different, showing (i) $\rightarrow$ (ii).

Suppose $T$ has infinitely many maximal paths. We define a sequence $x_0, x_1, \ldots$, a sequence $y_0, y_1, \ldots$, and a sequence $z_0, z_1, \ldots$ of elements of $T$ as follows. First, let $x_0 = \varepsilon$. Suppose we already defined $x_i$ in such a way that $T|_{x_i}$ has infinitely many maximal paths. We define $y_i$ as the (unique) minimal node $z \in T$ (with respect to $<_{\mathrm{pr}}$) being a (not necessarily strict) descendant of $x_i$ having at least two successors in $T$. Observe that $T|_{y_i}$ still has infinitely many maximal paths.

Having defined $y_i$, we define $x_{i+1}$ as a successor of $y_i$ in $T$ such that $T|_{x_{i+1}}$ still contains infinitely many maximal paths. (Since there are only a finite number of successors of $y_i$, such a node $x_{i+1}$ exists). Finally, let $z_i$ be an arbitrary successor of $y_i$ different from $x_{i+1}$.

It is easy to check that these sequences are well-defined and that the infinite set $\{z_0, z_1, \ldots\}$ contains pairwise incomparable nodes of $T$, showing (ii) $\rightarrow$ (i). $\square$

**Proposition 28.** *Suppose $G = (V, \Sigma, P, S, \mathcal{F})$ is an* MCFG *in normal form containing no reproductive nonterminals such that $L = L^\infty(G)$ contains scattered words only. Then the rank of each nonempty word in $L = L^\infty(G)$ is at most $|V|$. Moreover, for each symbol $X \in V \cup \Sigma$, the maximal rank $\#(X) = \max\{H(w) : w \in L^\infty(G, X)\}$ can be computed in polynomial time.*

**Proof.** If $P \subseteq \{S \to \varepsilon\}$, then $L$ contains no nonempty words and the statement trivially holds. So we can suppose $P \nsubseteq \{S \to \varepsilon\}$, hence each nonterminal is $+$-productive.

We claim that for any symbol $X \in V \cup \Sigma$, $\#(X)$ is the largest integer $n \geq 0$ for which there exist symbols $X_0, X_1, \ldots, X_n \in V \cup \Sigma$ such that $X_0 = X$ and for each $i \in [n]$, $X_{i-1} \rightsquigarrow^\infty X_i$. Since $G$ contains no reproductive nonterminals, $\rightsquigarrow^\infty$ is an irreflexive transitive relation, i.e. a strict partial ordering, so this value $n$ is well-defined for each $X \in V \cup \Sigma$.

Suppose $X_0, \ldots, X_m \in V \cup \Sigma$ are symbols with $X_{i-1} \rightsquigarrow^\infty X_i$ for each $i \in \{1, \ldots, m\}$. Since $G$ contains only $+$-productive nonterminals, applying Lemma 24 we get by induction on $m$ that $\#(X_0) \geq m$.

For a set $W \subseteq V$ of nonterminals and symbol $X \in W \cup \Sigma$, let $\#(W, X)$ denote the value $\max\{H(\mathrm{fr}(t)) : t \in \Delta(W, X)\}$. It suffices to show that whenever $\#(W, X) \geq k$ for some integer $k$, then there exist symbols $X_0, \ldots, X_k$ with $X = X_0$ and $X_{i-1} \rightsquigarrow^\infty X_i$ for each $i \in \{1, \ldots, k\}$. When either $k = 0$ or $X \in \Sigma$ is a terminal symbol, then this statement clearly holds. Suppose $X \in V$, thus $X \in W$ and $W \neq \emptyset$, $k > 0$, and suppose that we already shown the claim for each integer less than $k$, and for each set $W' \subsetneq W$ of nonterminals (for the integer $k$).

Let $t \in \Delta(W, X)$ be a derivation tree of $G$ such that $H(\mathrm{fr}(t)) \geq k$.

Let $T \subseteq \mathrm{dom}(t)$ be the set of those nodes of $t$ having a (not necessarily strict) descendant labeled $X$. Then $T$ is a sub-tree domain of $\mathrm{dom}(t)$. Since $X$ is not reproductive, there cannot be an infinite number of pairwise incomparable elements of $T$ (with respect to the prefix relation): if there was an infinite number of such nodes, there would also exist an infinite number of pairwise incomparable nodes of $t$ each labeled $X$ by Lemma 27, which would then contradict the non-reproductivity of $X$.

Thus, whenever $T' \subseteq T$ is a set of incomparable elements of $T$ (with respect to the prefix relation), then $T'$ is finite. Hence, $T$ is a finite union of some maximal paths $\pi_1, \ldots, \pi_r$.

Now let $T^+$ be the set of those successors of the nodes of $T$ that are not in $T$. Observe the following facts:

- The elements of $T^+$ are pairwise incomparable with respect to the prefix ordering.
- $\mathrm{fr}(t) = \prod_{x \in T^+} \mathrm{fr}(t|_x)$, where the ordering of elements of $T^+$ is the lexicographic ordering.
- For each $x \in T^+$, $t|_x$ is in $\Delta(W - \{X\}, t(x))$.

Since $T$ has a finite number of maximal paths, the order type of $T^+$ is a finite sum of order types, each being either finite, or one of $\omega$ and $-\omega$. To see this, suppose $T$ is the finite union of the maximal paths $\pi_1, \ldots, \pi_r$. We proceed by induction on $r$. If $r = 1$, then the order type of $T^+$ can be embedded into $\omega + (-\omega)$. Now let $r > 1$ and let $u$ be the longest common prefix of the paths $\pi_1, \ldots, \pi_r$. Then $T^+$ can be written as the ordered sum

$$\begin{aligned} T^+ = & \{u' \cdot i \in \mathrm{dom}(t) : \exists j > i, v\, u = u' \cdot j \cdot v\} + \\ & \left( \sum_{u \cdot i \in \mathrm{dom}(t)} L_i \right) + \\ & \{u' \cdot i \in \mathrm{dom}(t) : \exists j < i, v\, u = u' \cdot j \cdot v\}, \end{aligned}$$

where for each $i$ with $u \cdot i \in \mathrm{dom}(t)$, $L_i$ is either the singleton ordering $\{u \cdot i\}$ (if $u \cdot i \in \mathrm{dom}(t) - T$) or $(T|_{u \cdot i})^+$, where the subtree $(T|_{u \cdot i})$ is a union of $r_i < r$ maximal paths. Applying the induction hypothesis we get that the order type of $T^+$ is a finite sum of order types, each being finite or a finite sum of the order types $\omega$, $-\omega$ and the finite order types. The claim is proved.

Since $H(\mathrm{fr}(t)) \geq k$, one of the following cases holds:

- either $H(\mathrm{fr}(t|_x)) \geq k - 1$ for infinitely many nodes $x \in T^+$;
- or $H(\mathrm{fr}(t|_x)) = k$ for some node $x \in T^+$.

If $H(\mathrm{fr}(t|_x)) \geq k-1$ for infinitely many nodes $x \in T^+$, then there exists some symbol $Y \in V \cup \Sigma$ and an infinite set $T_Y \subseteq T^+$ with each $x \in T_Y$ labeled by $Y$ (thus $X \rightsquigarrow^\infty Y$) and $H(t|_x) \geq k - 1$ for each $x \in T_Y$. Applying the induction hypothesis on $k$ we get that there exist symbols $Y_0, \ldots, Y_{k-1} \in V \cup \Sigma$ such that $Y_0 = Y$ and $Y_{i-1} \rightsquigarrow^\infty Y_i$ for each $i \in \{1, \ldots, k\}$, thus by $X \rightsquigarrow^\infty Y$ we have $X \rightsquigarrow^\infty Y_0 \rightsquigarrow^\infty \cdots \rightsquigarrow^\infty Y_{k-1}$, showing the claim.

If $H(\mathrm{fr}(t|_x)) = k$ for some node $x \in T^+$, then (since $t|_x \in \Delta(W - \{X\}, t(x))$) applying the induction hypothesis on $W - \{X\} \subsetneq W$ we get there exist symbols $Y_0, \ldots, Y_k \in V \cup \Sigma$ with $Y_0 = t(x)$ and $Y_{i-1} \rightsquigarrow^\infty Y_i$ for each $i \in \{1, \ldots, k\}$. Now $Y_0$ is accessible from $X$ (since it occurs in a derivation tree whose root is labeled by $X$) and $Y_0 \rightsquigarrow^\infty Y_1$ (recall that $k > 0$), so that $X \rightsquigarrow^\infty Y_1$ and thus $X \rightsquigarrow^\infty Y_1 \rightsquigarrow^\infty \cdots \rightsquigarrow^\infty Y_k$, completing the proof. □

Corollary 25 and Proposition 28 immediately yield:

**Theorem 29.** *Suppose $L$ is an* MCFL *consisting only scattered words. Then one of the following cases holds:*

(i) *either there exists a finite bound $n$ such that each word in $L$ has rank at most $n$;*
(ii) *or for any countable ordinal $\alpha$, there exists a word in $L$ with rank at least $\alpha$.*

*Moreover, it can be decided in polynomial time whether* (i) *or* (ii) *holds, and if* (i) *holds, even the least such bound $n$ can be computed in polynomial time, if $L$ is given by an* MCFG *in normal form.*

**Theorem 30.** *It can be decided in polynomial time whether an* MCFL *given by an* MCFG *in normal form contains dense words only.*

**Proof.** Suppose $G = (V, \Sigma, P, S, \mathcal{F})$ is an MCFG in normal form. Then $L = L^\infty(G)$ contains a word which is *not* dense if one of the following conditions hold:

(i) $u \in L$ for some $u \in \Sigma \cup \{\epsilon\}$;
(ii) $S \Rightarrow^\infty \alpha ab\beta$ for some $\alpha, \beta \in (V \cup \Sigma)^\infty$ and $a, b \in \Sigma$.

By Proposition 16, (i) can be decided in polynomial time. Moreover, (ii) holds if and only if $S \Rightarrow^* \alpha' ab\beta'$ for some $\alpha', \beta' \in (V \cup \Sigma)^*$, thus it can also be decided in polynomial time. □

## 8. Closure properties

**Proposition 31.** *The class of* MCFL*s is effectively closed under substitution, i.e., when $L \subseteq \{a_1, \ldots, a_n\}^\infty$, $L_1, \ldots, L_n \subseteq \Sigma^\infty$ are* MCFL*s, each given by an* MCFG*, then an* MCFG *$G$ with $L^\infty(G) = L[a_1 \leftarrow L_1, \ldots, a_n \leftarrow L_n]$ can be given effectively.*

**Proof.** Let $L = L^\infty(G)$ where $G = (V, \Sigma, P, S, \mathcal{F})$, and for each $a \in \Sigma$, let $L_a = L^\infty(G_a)$, where $G_a = (V_a, \Delta, P_a, S_a, \mathcal{F}_a)$. Without loss of generality we may assume that the sets $V, V_a,\ a \in \Sigma$ are pairwise disjoint. Now let $P'$ be the set of productions obtained from the productions in $P$ by replacing each occurrence of each letter $a \in \Sigma$ with $S_a$. Then let

$$\begin{aligned}
\overline{V} &= V \cup \bigcup_{a\in\Sigma} V_a \\
\overline{P} &= P' \cup \bigcup_{a\in\Sigma} P_a \\
\overline{\mathcal{F}} &= \mathcal{F} \cup \bigcup_{a\in\Sigma} \mathcal{F}_a
\end{aligned}$$

and $\overline{G} = (\overline{V}, \Delta, \overline{P}, S, \overline{\mathcal{F}})$. The MCFG $\overline{G}$ generates the language $L[a \leftarrow L_a]_{a\in\Sigma}$. □

**Corollary 32.** *The class* MCFL *is closed under binary set union, concatenation, star, $\omega$-, $\eta$-, $\infty$-, and $(-\omega)$-power.*

**Proposition 33.** MCFL *is neither closed under intersection nor under complementation.*

**Proof.** By Corollary 17, a language of finite words is an MCFL if and only if it is context-free. Since the class of context-free languages is not closed under intersection, the class of MCFLs is also not closed under intersection, and hence also not closed under complementation (since it is closed under binary union). □

Recall the definition of the languages $\mathrm{Pre}(L)$, $\mathrm{Suf}(L)$, $\mathrm{In}(L)$ and $\mathrm{Sub}(L)$.

**Proposition 34.** *If $L$ is an* MCFL*, then $L^r$, $\mathrm{Pre}(L)$, $\mathrm{Suf}(L)$, $\mathrm{In}(L)$ and $\mathrm{Sub}(L)$ are also* MCFL*s.*

**Proof.** Suppose that $L$ is an MCFL generated by the MCFG $G = (N, \Sigma, P, S, \mathcal{F})$. It is clear that $L^r$ is generated by the MCFG $G^r = (N, \Sigma, P^r, S, \mathcal{F})$ where $P^r = \{X \to p^r : X \to p \in P\}$.

Regarding $\mathrm{Pre}(L)$, let $\overline{N} = \{\overline{X} : X \in N\}$ and $\overline{F} = \{\overline{X} : X \in F\}$. Then consider the grammar $\mathrm{Pre}(G) = (N \cup \overline{N}, \Sigma, P \cup \overline{P}, \overline{S}, \mathcal{F} \cup \overline{\mathcal{F}})$, where

$$\begin{aligned}
\overline{P} &= \{\overline{X} \to p\overline{Y} : X, Y \in N,\ \exists q\, X \to pYq \in P\} \\
&\quad \cup \{\overline{X} \to pa : X \in N, a \in \Sigma,\ \exists q\, X \to paq \in P\} \\
&\quad \cup \{\overline{S} \to \varepsilon\},\ \text{and} \\
\overline{\mathcal{F}} &= \{\overline{F} : F \in \mathcal{F}\} \text{ where } \overline{F} = \{\overline{X} : X \in F\}.
\end{aligned}$$

If $G$ is in normal form with $P \not\subseteq \{S \to \varepsilon\}$, then $G' = \mathrm{Pre}(G)$ generates $\mathrm{Pre}(L)$. To see this, consider a derivation tree $t$ over the grammar $G$ whose root symbol is $S$ and whose frontier word $u$ is in $\Sigma^\infty$. If $v$ is a prefix of $u$ then we can partition the set of leaves into two disjoint sets $K$ and $R$ such that $K$ is closed below and $R$ is closed above with respect to the lexicographic order and such that $v$ is isomorphic to the word determined by $K$. If $K$ is empty then $v$ is the empty word and since $\overline{S} \to \varepsilon$ is a production of $G'$ we have $v = \varepsilon \in L^\infty(G')$. Assume now that $K$ is not empty. Using $t$, we will construct a derivation tree for $v$ over the grammar $G'$. To this end, let us relabel the root by $\overline{S}$. Then suppose that we have relabeled a node $x$ originally labeled $X \in N$ by $\overline{X}$ such that every node in $K$ is either lexicographically less than $x$ or belongs to $t|_x$, moreover, $t|_x$ contains at least one leaf in $K$. Consider the successors $x_1, \ldots, x_m$ of $x$. There is a largest integer $i$ such that the subtree $t|_{x_i}$ rooted at $x_i$ contains a leaf in $K$. If $x_i$ is labeled in $\Sigma$, or $i = 1$ and $x_i$ is labeled $\varepsilon$, then $x_i$ is the lexicographically greatest element of $K$. A derivation tree may be obtained from the relabeled tree by removing all nodes lexicographically greater than $x_i$. If $x_i$ is labeled by a nonterminal $Y$ then we relabel it $\overline{Y}$ and continue the process. If the process does not stop, then the nodes which are relabeled form an infinite path $\pi$ and a leaf belongs to $K$ if and only if it is lexicographically less than $\pi$. A derivation tree of $v$ over $G'$ can be obtained by removing all nodes lexicographically greater than $\pi$.

Suppose now that $t$ is a derivation tree over $G'$ with root symbol $\overline{S}$ and frontier word $v$ in $\Sigma^\infty$. Then the inner nodes of $t$ on the rightmost complete path $\pi$ are labeled by nonterminals in $\overline{N}$ and all other inner nodes are labeled in $N$. Suppose that $x$ is an inner node lying on $\pi$ labeled $\overline{X}$. Let $x_1, \ldots, x_i$ denote the successors of $x$ labeled $p_1, \ldots, p_i$, respectively. If $p_i = \overline{Y}$ is in $\overline{N}$ then there are some $q_1, \ldots, q_j \in N \cup \Sigma$ such that $X \to p_1 \ldots p_{i-1} Y q_1 \ldots q_j$ is a production of $G$. In this case let us add $j$ new successors of $x$ to the tree, labeled $q_1, \ldots, q_j$, respectively. If $p_i$ is a terminal or $i = 1$ and $p_i = \varepsilon$, then $x_i$ is the last node of $\pi$. Moreover, there exist $q_1, \ldots, q_j \in N \cup \Sigma$ such that $X \to p_1 \ldots p_i q_1 \ldots q_j$ is in $P$. We add $j$ new successors of $x$ to the tree, labeled $q_1, \ldots, q_j$, respectively. Replacing each node label $\overline{X}$ with $X$, the tree constructed in this way is a derivation tree over $G$ whose frontier word is of the form $vq$ for some $q \in (N \cup \Sigma)^\infty$. Since $G$ contains only productive nonterminals, this tree can be completed to a derivation tree whose frontier word $u$ is in $\Sigma^\infty$. It is clear that $v$ is a prefix of $u$.

Since $\mathrm{Suf}(L) = (\mathrm{Pre}(L^r))^r$ and $\mathrm{In}(L) = \mathrm{Suf}(\mathrm{Pre}(L))$, it follows now that $\mathrm{Suf}(L)$ and $\mathrm{In}(L)$ are also MCFLs.

Last, we prove that $\mathrm{Sub}(L)$ is an MCFL. For this reason, without loss of generality we may assume that whenever a terminal letter $a$ occurs on the right side of a production, then the production is of the form $X \to a$. If $G$ satisfies this condition, then a grammar generating $\mathrm{Sub}(L)$ is obtained by adding all productions $X \to \varepsilon$ to the set $P$ whenever $X \to a$ is in $P$ for some $a \in \Sigma$. $\square$

## 9. Conclusion, open questions

We have defined Müller context-free grammars (MCFGs) generating languages of countable words, called MCFLs. The class of MCFLs is clearly closed under substitution and thus enjoys good closure properties. We have studied several decision problems for MCFLs, mainly motivated by order theoretic properties, and in each case we have found a polynomial time algorithm for MCFGs in normal form. The transformation of an arbitrary grammar into normal form requires polynomial space, but the size of the grammar produced by the algorithm is polynomial in the size of the input grammar. Among the decision problems, we showed that it is decidable in polynomial time whether an MCFG in normal form generates a language of well-ordered, or scattered, or dense words. We have established a limitedness property: If an MCFL contains only scattered words, then either the rank of each word of the language is bounded by a fixed integer $n$, or for each countable ordinal $\alpha$ there is a word in the language of rank at least $\alpha$. Moreover, we have shown that is decidable which of the two cases applies.

In an earlier paper we studied Büchi context-free languages, or BCFLs. While every BCFL is an MCFL, there exists an MCFL of scattered, or even well-ordered words that is not a BCFL. It remains for future research to answer the question whether there is a MCFL consisting of dense words that is not a BCFL. On the other hand, it not difficult to show that every MCFL consisting of finite or $\omega$-words is a BCFL. By a result in [10], it then follows that an $\omega$-language is an MCFL if and only if it is context-free in the sense of Cohen and Gold [8].

The equality problem for BCFLs is undecidable, [10], thus it is also undecidable for MCFLs. We have not yet studied the question of deciding whether an MCFG generates a BCFL. Also, it would be interesting to know whether there is an MCFL of scattered words of rank bounded by an integer $n$ that is not a BCFL.

## References

[1] N. Bedon, Finite automata and ordinals, Theoret. Comput. Sci. 156 (1996) 119–144.
[2] A. Bès, O. Carton, A Kleene theorem for languages of words indexed by linear orderings, in: DLT'2005, in: LNCS, vol. 3572, Springer, 2005, pp. 158–167.
[3] S.L. Bloom, Z. Ésik, Deciding whether the frontier of a regular tree is scattered, Fund. Inform. 55 (2003) 1–21.
[4] L. Boasson, Context-free sets of infinite words, in: Theoretical Computer Science (Fourth GI Conf., Aachen, 1979), in: LNCS, vol. 67, Springer, 1979, pp. 1–9.
[5] V. Bruyère, O. Carton, Automata on linear orderings, J. Comput. System Sci. 73 (2007) 1–24.
[6] J.R. Büchi, The monadic second order theory of $\omega_1$, in: Decidable Theories, II, in: Lecture Notes in Math., vol. 328, Springer, 1973, pp. 1–127.
[7] Y. Choueka, Finite automata, definable sets, and regular expressions over $\omega^n$-tapes, J. Comput. System Sci. 17 (1) (1978) 81–97.
[8] R.S. Cohen, A.Y. Gold, Theory of $\omega$-languages, parts one and two, J. Comput. Syst. Sci. 15 (1977) 169–208.
[9] B. Courcelle, Frontiers of infinite trees, RAIRO Theor. Inf. 12 (1978) 319–337.
[10] Z. Ésik, Sz. Iván, Context-free languages of countable words, in: ICTAC 2009, in: LNCS, vol. 5684, Springer, 2009, pp. 185–199.
[11] F. Gécseg, M. Steinby, Tree automata, Akadémiai Kiadó, Budapest, 1984.
[12] S. Heilbrunner, An algorithm for the solution of fixed-point equations for infinite words, RAIRO Theor. Inf. 14 (1980) 131–141.
[13] P. Hunter, A. Dawar, Complexity bounds for regular games, in: Proc. of Mathematical Foundations of Computer Science (MFCS) 2005, in: LNCS, vol. 3618, 2005, pp. 495–506.
[14] A. Montalbán, On the equimorphism types of linear orderings, Bull. Symbolic Logic 13 (1) (2007) 71–99.
[15] M. Nivat, Sur les ensembles de mots infinis engendrés par une grammaire algébrique, RAIRO Inform. Théor. 12 (3) (1978) 259–278 (in French).
[16] J.G. Rosenstein, Linear Orderings, Academic Press, 1982.
[17] J. Wojciechowski, Classes of transfinite sequences accepted by finite automata, Fund. Inform. 7 (1984) 191–223.
[18] J. Wojciechowski, Finite automata on transfinite sequences and regular expressions, Fund. Inform. 8 (1985) 379–396.